

# 基于变分贝叶斯的纤维丛元学习算法



刘洋 李凡长

苏州大学计算机科学与技术学院 江苏 苏州 215006

(20185227047@stu.suda.edu.cn)

**摘要** 以神经网络为基础的深度学习在大量领域取得优异成果,但其难以处理相似或未经训练的任务。深度学习在对新任务的学习和适应过程中存在困难,且对训练样本规模要求很高,造成泛化性和扩展性不佳的问题。元学习是一种新的学习框架,旨在解决传统学习方法难以解决的快速学习和适应新任务的问题。针对图像分类的元学习问题,文中提出了一种基于贝叶斯理论的纤维丛元学习算法(Fiber Bundle Meta-learning Algorithm, FBBML)。首先通过卷积神经网络提取支持数据集的图片信息,以得到图片的表示。然后构建数据特征的流形结构和数据特征到标签的纤维丛。最后输入查询集选取当前新任务的流形截面,从而获得适合新任务的纤维,得到图片的正确标签。实验结果表明,基于所提算法实现的模型(FBBML)在公共数据集(mini-ImageNet)上相比标准四层卷积神经网络的模型取得了最佳的准确率性能。同时将纤维丛理论引入元学习,使得算法本身具备更高的可解释性。

**关键词:** 贝叶斯;流形;纤维丛;分类;卷积神经网络;元学习

中图分类号 TP301.6

## Fiber Bundle Meta-learning Algorithm Based on Variational Bayes

LIU Yang and LI Fan-zhang

School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China

**Abstract** Deep learning based on neural network has achieved excellent results in a large number of fields, but it is difficult to deal with similar or untrained tasks, and it is difficult to learn and adapt to new tasks. Moreover, it requires a high scale of training samples, resulting in its poor generalization and expansion. Meta learning is a new learning framework, which aims to solve the problem that traditional learning methods can't solve fast learning and adapt to new tasks. Aiming at the meta learning problem of image classification, a novel fiber bundle meta learning algorithm based on Bayesian theory is proposed. Firstly, the convolution neural network is used to extract the image information supporting the dataset, and the image representation is obtained. Then the manifold structure of data features and the fiber bundle of data features are constructed. The input query set selects the manifold section of the current new task to obtain the fiber suitable for the new task, so as to get the correct label of the image. Experimental results show that the model based on the proposed algorithm (FBBML) achieves the best accuracy performance compared with the standard four-layer convolutional neural network model on the common data set (mini-ImageNet). At the same time, the fiber bundle theory is introduced into meta learning, which makes the algorithm more interpretable.

**Keywords** Bayesian, Manifold, Fiber bundle, Classification, Convolution neural network, Meta learning

## 1 引言

机器学习往往围绕一个具体的任务展开。但从生物学的角度来看,生物一生中并不是只解决一个任务。生物体具有一种特别的能力,通过对已经观察的任务学习,能够自然而然地解决未观察的任务。元学习<sup>[1]</sup>本质上就是模拟生物的这一特殊能力,通过学习已经观测的任务来解决未观测的任务。

传统的机器学习模式是获取特定任务的大型数据集,用数据集训练模型。这与生物利用以往经验,仅仅通过少量

新的未观测样本就迅速完成学习的情况相差甚远。

在多数现实场景中,可能无法获得足量的数据供学习系统进行学习,因此学习系统只能学习少量数据,如部分医疗场景。在这种场景下,学习系统需要快速地适应新的任务。针对上述小样本问题,具有学习先验知识子系统的元学习成为当前小样本学习的优先选择。

深度学习<sup>[2]</sup>是一种黑盒的端到端的学习模式,深度学习能自行学习数据的规则,并且能得到较为理想的效果,从而广泛地应用于各个领域。但深度学习存在着可解释性差、泛化

到稿日期:2020-11-16 返修日期:2021-03-10 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家重点研发计划(2018YFA0701700,2018YFA0701701);国家自然科学基金(61902269)

This work was supported by the National Key R & D Program of China (2018YFA0701700, 2018YFA0701701) and National Natural Science Foundation of China (61902269).

通信作者:李凡长(lfzh@suda.edu.cn)

性差等根本问题。此外深度学习还存在难以利用先前任务信息的问题。当前的深度学习大部分情况下只能从头开始训练,使用 Fine-tune 来学习新任务,其效果往往不能达到要求,而元学习在深度学习上的研究就是如何让神经网络更好地利用以往的知识,使得元学习方法能根据新任务调整学习器。

从目标方面来看,元学习和迁移学习<sup>[3]</sup>非常相似,都是考虑如何提高学习器泛化能力的研究。但迁移学习更多地考虑如何将泛化能力从一个任务迁移到其他任务。总体来说它是考虑一个任务的扩展性。元学习是从整体考虑,考虑如何在不同的任务之间采集其解决方案的共同规律,从而形成泛化性更强的先验知识。因此,元学习在采样数据的同时也需要采样任务,通过学习任务之间的相关信息来帮助模型更好地泛化。

本文的主要贡献如下:

- (1) 将纤维丛理论引入元学习的表示,提出了基于纤维丛理论的元学习算法框架,提高了元学习的可解释性。
- (2) 基于上述元学习框架,结合贝叶斯理论,提出了一个纤维丛贝叶斯元学习模型的具体实例。
- (3) 在小样本数据集上进行了大量实验,并且取得了较为理想的效果。

## 2 相关工作

元学习不是一个新的概念。在机器学习领域成熟之前已有相关概念。文献[4]提出了“元分析”这一概念,对大量的分析结果进行统计分析,这是一种二次分析的方法。文献[5]首次提出了元学习这一概念,将元学习描述为“学习者意识到并越来越多地控制他们已经内化的感知、探究、学习和成长习惯的过程”,将元学习作为在假设、结构、变化、过程和发展这5个方向的综合体,并阐述了相关基本原则。随后元学习这一概念慢慢在机器学习领域发展。文献[6]提出的元学习是一种整合多种学习过程的技术,利用元学习的策略组合多个不同算法设计的分类器,其整体的准确度优于任何个别的学习算法。文献[7]提出了基于元学习的决策树框架。文献[8]认为元学习是通过积累元知识动态地利用经验来改善偏差的一种学习算法。

此后,元学习在深度学习方面的发展相对缓慢。直到文献[9]提出 MAML (Model-Agnostic Meta-Learning) 方法,提出利用不同任务的梯度信息作为先验知识进行学习更新。这使得元学习重新成为研究热点之一。

一般来说,元学习系统应具有以下3个特点:1)元学习系统利用以前学习中提取的元知识来获得经验;2)元知识来自单个数据集或不同领域;3)动态选择学习偏差。

流形学习最早是一种非线性维数约简方法,文献[10]从认知上讨论了流形学习,最早使用了 Manifold Learning 这一学术名词,强调认知过程的完整性。文献[11]提出了流形学习的通用定义:设  $Y \subset R^D$  是一个低维流形,  $f: Y \rightarrow R^d$  是一个光滑嵌入,其中  $D > d$ 。数据集  $\{y_i\}$  是随机生成的,且经过  $f$  映射为观察空间的数据  $\{x_i = f(y_i)\}$ 。流形学习就是在给定观察样本集  $\{x_i\}$  的条件下重构  $f$  和  $\{y_i\}$ 。从数学上来看,流形学习问题的本质是一类寻找非线性的维数约简方法的学习

问题。流形学习至今已有了多种发展方向,按照数据空间的分析方法不同可分为神经网络法、变分法、互信息法、主流形分析法、谱方法等几个研究方向。本文将纤维丛扩展到元学习中,故采用神经网络分析方法。

## 3 基本概念

本节将给出本文所需的相关基本定义。

**定义 1** (纤维丛 (Fiber bundle)<sup>[12]</sup>) 一个纤维丛  $(E, M, \pi)$  由一个底流形  $M$  和一个全空间  $E$  组成,在全空间  $E$  和底流形  $M$  之间有一个投影变换  $\pi: E \rightarrow M$ , 对应的逆变换在  $M$  上的每一点  $x$  处得到纤维  $F_x, \pi^{-1}: x \rightarrow F_x$ , 纤维  $F_x$  称为通过  $x$  点的纤维,如果纤维  $F_x$  是  $k$  维的向量空间,则称纤维丛  $(E, M, \pi)$  为矢丛。

纤维丛是流形向乘积的推广。矢丛  $E = \bigcup_{x \in M} F_x$  是底流形  $M$  上所有点的纤维的并集,是以流形  $M$  为参数的向量空间族,局部可表示为参数空间  $U$  与向量空间  $F$  的拓扑积  $U \times F$  但是丛的整体不能表示为平均拓扑积,向量丛  $E$ 。是积流形与纤维粘附的结果,并且要求纤维上的线性关系保持不变。光纤束的局部拓扑结构简单,但整体的拓扑结构取决于如何在重叠区平滑地结合光纤。

**定义 2** 流形上点的切空间的并集为:

$$T(M) = \bigcup_p T_p(M) \quad (1)$$

称为流形  $M$  的切丛<sup>[12]</sup>。切丛  $T(M)$  局部是直积流形,但整体不一定是直积流形。

**定义 3** 向量场  $X(x)$  又称为切丛  $T(M)$  的一个截面,是在流形  $M$  上每点  $p$  选出一个切向量  $X_p \in T_p(M)$ , 当选局部坐标系后,切向量场  $X(x)$  可表示为:

$$X(x) = \sum_{j=1}^n \xi^j(x) \partial_j \quad (2)$$

如果函数  $\xi^j(x)$  连续可微,则称向量场  $X(x)$  连续可微。

**定义 4** 设在纤维丛有  $\pi: E \rightarrow M$  是光滑流形  $M$  上的矢丛,  $U \subset M$  为开集。若有光滑映射  $s: U \rightarrow E$  使得:

$$\pi \circ s = id: U \rightarrow U \quad (3)$$

则称  $s$  为矢丛  $E$  定义在  $U$  上的一个光滑截面<sup>[12]</sup>。当  $U = M$  时,则称  $s$  为矢丛  $E$  的一个整个光滑截面。矢丛  $E$  的光滑截面的集合记为  $\Gamma(E)$ 。

**定义 5** 矢丛  $E$  上的联络<sup>[12]</sup> 是一个映射:

$$D: \Gamma(E) \rightarrow \Gamma(T^*(M) \otimes E) \quad (4)$$

映射  $D$  满足下列条件:

(1) 对于  $\forall s_1, s_2 \in \Gamma(E)$  有:

$$D(s_1 + s_2) = Ds_1 + Ds_2 \quad (5)$$

(2) 对于  $s \in \Gamma(E)$  及  $\forall \alpha \in C^\infty(M)$ , 有:

$$D(\alpha s) = d\alpha \otimes s + \alpha Ds \quad (6)$$

切丛  $T(M)$  上的联络叫做流形  $M$  的仿射联络。

**定义 6** 设  $C: u' = u'(t)$  是底流形  $M$  上的一条参数曲线,  $X(t)$  是定义在  $C$  上的切向量场<sup>[12]</sup>:

$$X(t) = x^i(t) \left( \frac{\partial}{\partial u^i} \right)_{C(t)} \quad (7)$$

## 4 纤维丛元学习算法

### 4.1 小样本分类问题设置

小样本学习问题 (Few-shot)<sup>[13]</sup> 以任务 (Task) 为基本

单元。一个任务数据集(见图1)分为支持集(Support Set)和查询集(Query Set)。支持集包括  $N$  个分类标签,记为  $S = \{C_i | i=1 \dots N\}$ ,每个类别有  $k$  个样本,记为  $D_s = \{(x_i^s, y_i^s) | i=1 \dots k\}$ ,共有  $N * k$  个样本。查询集包括  $q$  个未分类样本,记为  $D_q = \{x_i^q | i=1 \dots q\}$ 。小样本分类问题是在少量支持集下对查询集进行分类,一般来说从元学习研究角度,小样本学习问题的  $k$  小于 10,查询集和支持集的标签无交集。上述小样本问题描述为 N-way k-shot 问题。

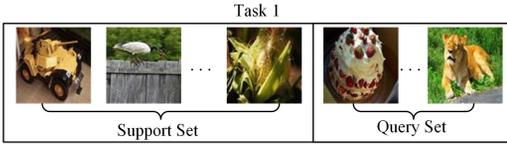


图1 元学习任务示意图

Fig. 1 Schematic diagram of meta-learning task

## 4.2 通用元学习

一般来说元学习是通过元训练阶段(见图2)得到先验知识。在元测试阶段(见图3)利用先验知识快速学习和适应新的任务。先验知识包括泛化性较强的网络初始化参数、元特征表示等。元学习的核心问题是如何表示和学习先验知识,从而利用先验知识解决小样本问题。

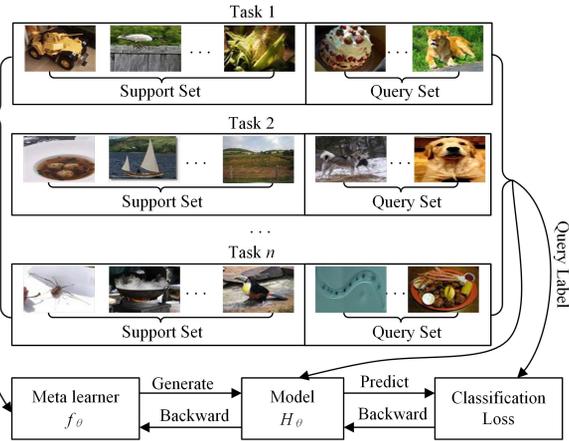


图2 元学习元训练阶段示意图

Fig. 2 Schematic diagram of meta-training stage

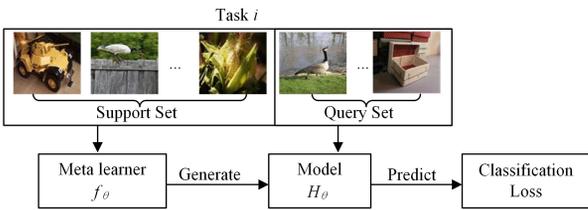


图3 元学习元测试阶段示意图

Fig. 3 Schematic diagram of meta-test stage

元学习的元训练阶段的一般流程如下:将数据集分为支持集和查询集,并采样任务;利用支持集生成元学习器,即得到先验知识;输入查询集得到查询集的预测值;利用查询集的预测值来更新模型参数和先验信息。

元测试阶段的基本流程如下:利用学习好的先验知识和元学习的测试训练数据继续优化模型,然后输入元测试的查询集得到预测值。

## 4.3 纤维丛元学习算法框架

纤维丛学习<sup>[14]</sup>是一种面向高维非线性数据的降维方法,通过找到高维空间中的低维流形,求出相应嵌入映射,以达到数据约简的目的。为了将纤维丛理论引入元学习,我们采用流形学习<sup>[15]</sup>的通用假设:已观察到的数据实际上是从低维流形映射到高维空间。由于数据内部特性的限制,一些高维数据会产生信息的冗余,只有相对较低的维度才能被唯一地表示出来。因此,任务的信息可以在一个较低的本真维度<sup>[16]</sup>下进行表示。新任务可以通过元学习进行分类。在获得数据流形的低维表示之后,相同类的对象在得到相同的表示时将得到相同的或者相近的低维坐标。元学习的先验知识是低维坐标系的学习和纤维丛映射的逼近。元学习通过元训练获得数据低维空间  $E$  坐标系和投影映射函数,空间坐标系是由支持集的实例决定的。元测试则是在本真维度下距离分布的度量。尽管支持集和查询集的类别不同,但由于本真维度下类别相近的距离更近,从而保证了小样本的查询集不会因为网络学习能力强而坍塌到一起产生过拟合的问题。因此,元学习可以快速适应新的任务,从而提高其性能。

基于上述理论,本文给出纤维丛元学习算法的通用框架。首先定义算法相关符号:元训练阶段支持集为  $D_{tr} = \{(x_i^{tr}, y_i^{tr}) | i=1 \dots k * N\}$ ,其中  $k$  为 shot 数,  $N$  为类别数,即对应小样本学习的 N-way k-shot 问题。元训练阶段的查询集为  $D_{te} = \{(x_i^{te}, y_i^{te}) | i=1 \dots q\}$ 。元测试阶段数据集同样分为支持集  $D_{tr}^i = \{(x_i^{tr}, y_i^{tr}) | i=1 \dots k * N\}$  和查询集  $D_{te}^i = \{(x_i^{te}) | i=1 \dots q\}$ 。根据上述小样本学习问题的定义,给出纤维丛元学习通用算法框架,包括纤维丛元学习元训练算法(见算法1)和纤维丛元学习元测试算法(见算法2)。

### 算法1 纤维丛元学习元训练算法

输入:元训练支持集  $D_{tr}^i$  和查询集  $D_{te}^i$

输出:元参数  $\theta$  和网络参数  $\phi$

1. 初始化元参数  $\theta$  和网络参数  $\phi$
2. 利用支持集构造纤维丛的投影映射  $\pi$
3.  $E^* \leftarrow D_{tr}^i, M^* \leftarrow \{\theta_i | \theta_i \in U_{\theta_i}\}$ , 得到纤维丛  $(E^*, M^*, \pi)$
4. for  $T_i \in E^*$  do

支持集类别本真表示  $\theta_i \leftarrow \pi((x_i^{str}, y_i^{str}))$

度量  $l = d(\theta, x_i^{tr})$

更新参数  $\phi_{i+1} \leftarrow L(l, y_i^{tr}, \phi_i, \theta_i)$

5. 输出元参数  $\theta, \pi, \phi$ 。

### 算法2 纤维丛元学习元测试算法

输入:元测试支持集  $D_{tr}^i$  和查询集  $D_{te}^i$ , 元参数  $(\theta^*, \phi^*, \pi^*)$

输出:元测试查询集标签预测值  $Y_{te}^q$

1.  $\theta \leftarrow \theta^*, \phi \leftarrow \phi^*, \pi \leftarrow \pi^*$
2.  $E^* \leftarrow D_{tr}^i, M^* \leftarrow \{\theta_i | \theta_i \in U_{\theta_i}\}$ , 得到纤维丛  $(E^*, M^*, \pi)$
3. for  $T_i \in E^*$  do

支持集类别本真表示  $\theta_i \leftarrow \pi((x_i^{str}, y_i^{str}))$

度量  $l = d(\theta, x_i^{te})$

预测值  $y_{te}^q \leftarrow f(l, \phi_i, \theta_i)$

4. 输出:  $Y_{te}^q = \{y_{te}^q\}$

算法1中,  $E^*$  为底流形  $M$  的子集,  $M^*$  为本真全空间  $E$  的子集。  $d$  为度量函数,  $\theta_i$  为第  $i$  个任务支持集类别的本真表示,例如采样编号为  $i$  的任务,其支持集共有 4 个类别,  $\theta_i = \{\theta_j^i | j=1 \text{ to } 4\}$ 。在此框架下元训练阶段的任务为学习先验

数据的流形结构和纤维丛的投影映射。通过投影映射,获得类别的低维表示。根据支持集类别的低维表示,构建低维流形坐标系,以坐标系作为辅助来进行查询集的分类,从而有效地利用先验信息。

此框架下的元测试算法则是利用训练阶段获得模型和先验信息,对学习系统没有学习过的类别进行快速学习和先验预测,这样可同时从支持集获得流形的局部坐标,可以较为理想地对未观察任务进行分类和预测,从而使得模型有较高的泛化能力。

### 5 贝叶斯纤维丛元学习算法的实现

本节将详细介绍基于上述纤维丛元学习算法框架和变分贝叶斯理论的元学习模型实例 FBBML,包括算法中的网络结构,以及整体的算法流程。

由上述纤维丛元学习框架可知,基于纤维丛元学习方法的关键问题是如何对纤维丛进行建模和选择重塑的本真维度。为解决第一个问题,本文借鉴变分自编码器<sup>[17]</sup>的思想,通过对支持集进行编码,得到二维的均值和方差,再通过采样得到本真维度的生成样本  $\theta$ ,同时在采样部分本模型也使用重参数技巧。但是与变分贝叶斯方法不同的是,本文方法生成的是支持集类别的均值和方差,并不是生成每个样本的均值和方差。

选择重塑的本真维度的关键在于如何利用先验信息,即利用变分自编码器相似的部分得到纤维丛,然后通过重参数采样获得本真维度类表示。通过类别的表示,可以构建流形的局部坐标,然后在此坐标下对未见类进行分类。

本文方法的示意图如图 4 所示,模型整体分为 3 个部分,分别是特征提取器、编码器和分类器。通过特征提取器提取支持集和查询集的特征,然后编码器得到查询集的低维表示,最后分类器对查询集进行分类。

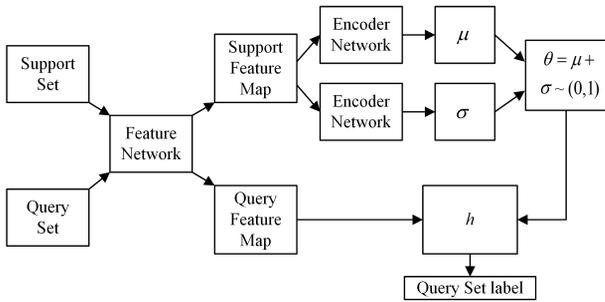


图 4 FBBML 的示意图

Fig. 4 Schematic diagram of FBBML

#### 5.1 FBBML 的网络结构

本模型网络主要分为 3 部分,其中特征提取部分采用标准卷积神经网络<sup>[18]</sup>,步长为(1,1),卷积核大小为 $[3 \times 3]$ ,filters 为 64,每层对应一层池化层,池化层窗口大小为 $[2 \times 2]$ ,步长为 2。在后续实验中,为减少特征提取对方法的影响,特征提取器均采用四层卷积神经网络,因此实验部分能较为公平地比较各个方法的性能。在编码器阶段,采用三层全连接网络,激活函数为 relu,分别对  $\mu$  和  $\sigma$  进行编码。

#### 5.2 FBBML 概率推导

本文采用变分编码结构来对纤维丛进行建模,变分编码结构通过变分法进行分布近似,元学习是基于多任务下的

一种学习方法,多任务学习<sup>[19]</sup>概率图如图 5 所示。

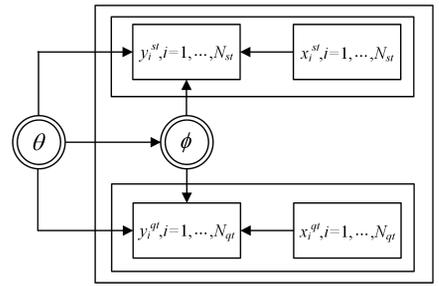


图 5 多任务概率图

Fig. 5 Directed graphical model for multi-task learning

图 5 中,  $\phi$  是网络参数,  $\theta$  是共享参数,  $x_s^t$  是支持集样本输入,  $y_s^t$  是支持集的标签输出,  $x_q^t$  是查询集样本输入,  $y_q^t$  是查询集标签输出。此处输入和输出包含元测试和元训练两个阶段。

由图 5 可以得到,对于任务  $t$ ,多任务学习的联合概率为:

$$p(\{Y^t, \phi^t\}_{t=1}^T | \{X^t\}_{t=1}^T, \theta) = \prod_{t=1}^T p(\phi^t | \theta) \prod_{i=1}^{N_s} p(y_i^s | x_i^s, \phi^t, \theta) \prod_j^Q p(\tilde{y}_j^t | \tilde{x}_j^t, \phi^t, \theta) \quad (8)$$

由式(8)可知,FBBML 的目标是求得未见类的后验概率:

$$p(\tilde{y}^t | \tilde{x}^t, \theta) = \int p(\tilde{y}^t | \tilde{x}^t, \phi^t, \theta) p(\phi^t | D^t) d\phi^t \quad (9)$$

为得到后验分布,引入一个近似分布  $q(\tilde{y}^t | \tilde{x}^t)$  来近似后验分布,如式(10)所示:

$$q(\tilde{y}^t | \tilde{x}^t, \theta) = \int p(\tilde{y}^t | \tilde{x}^t, \phi^t, \theta) q(\phi^t | D^t, \theta) d\phi^t \quad (10)$$

上述概率问题转化为优化式(11):

$$\arg \min E(p || q) = \arg \max E(\log \int p(\tilde{y} | \phi) q(\phi | D) d\phi) \quad (11)$$

通过蒙特卡洛近似<sup>[20]</sup>,可以得到 FBBML 训练时的损失函数为:

$$\mathcal{L}(\theta, \phi) = \frac{1}{TQ} \sum \log \frac{1}{L} \sum_{l=1}^L p(\tilde{y}_l^t | \tilde{x}_l^t, \phi_l^t, \theta) \quad (12)$$

其中,  $\phi_l^t \sim q(\phi | D^t, \theta)$ ,  $L$  为采样总数。

因为 FBBML 对类别进行编码,所以在近似分布计算上与一般变分自编码器有所不同,设全连接编码器为  $h$ ,则近似分布如式(13)所示:

$$q(\phi | D, \theta) = \prod_{c=1}^C q(\phi | \{h(x_n^c)\}_{n=1}^{K_c}, \theta) \quad (13)$$

其中,  $C$  表示已观察类也就是支持集类别数,  $K_c$  为每个类支持集已观察类  $c$  的样本总数。

### 6 实验结果与分析

实验使用 Omniglot<sup>[21]</sup>和 mini-ImageNet<sup>[22]</sup>数据集。Omniglot 数据集包含来自 50 个不同字母的 1623 个不同手写字符。每一个字符都由 20 个人通过亚马逊的 Mechanical Turk 在线绘制。Omniglot 相当于 1623 个类,每类 20 个样本。mini-ImageNet 抽样于 ImageNet 数据集,ImageNet 是常用的大型视觉数据集,ImageNet 为超过 1400 万张图像进行了注释,而且为其中 100 万张图像提供了边框。ImageNet 包含 2 万多个类别,每个类别均有不少于 500 张的图像。ImageNet 是一个非常庞大的视觉数据集,因此为研究小样本问题,

2016年 google DeepMind 团队 Vinyals 等在 ImageNet 的基础上采样选择提取出 mini-ImageNet 数据集。mini-ImageNet 包含 100 个类别,每个类别中包含 600 个样本数据。本文实验将其中 64 个类别数据作为训练集,16 个类别数据作为验证集,20 个类别数据作为测试集。

实验环境采用 TensorFlow-gpu rl.1.14.0 作为深度学习模型的框架。实验硬件环境如下:CPU 为 Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20 GHz,内存大小 250 GB, GPU 为 Ge-

Force RTX 2080Ti,显存大小 11 GB。

## 6.1 模型精度

本实验总共运行 12 种方法,每种方法在 Omniglot 数据集进行了 5-way 1-shot, 5-way 5-shot, 20-way 1-shot, 20-way 5-shot 实验;在 mini-ImageNet 数据集上进行了 5-way 1-shot 和 5-way 5-shot 实验。为了避免特征提取对模型的影响,对比方法的特征提取部分均采用 4 层卷积神经网络,各方法的准确率如表 1 所列。

表 1 Omniglot, mini-ImageNet 数据集上不同元学习方法准确率  
Table 1 Accuracy for different meta-learning methods on Omniglot and mini-ImageNet

(单位: %)

Method	Accuracy					
	Omniglot				mini-ImageNet	
	5-way accuracy		20-way accuracy		5-way accuracy	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Matching Nets	98.1	98.9	93.8	98.5	46.6	60.0
MAML	98.7±0.4	99.9±0.1	95.8±0.3	98.9±0.2	48.7±1.84	63.11±0.92
Prototypical Nets	97.4	99.3	95.4	98.7	46.61±0.78	65.77±0.70
mAP-SSVM	98.6	99.6	95.2	98.6	50.32±0.80	63.94±0.72
mAP-DLM	98.8	99.6	95.4	98.6	50.28±0.8	63.7±0.7
Meta-SGD	99.53±0.26	<b>99.93±0.09</b>	95.93±0.38	98.97±0.19	50.47±1.89	64.3±0.94
SNAIL	99.07±0.16	99.78±0.09		<b>99.36±0.18</b>	45.1	55.2
Relation Net	99.6±0.2	99.8±0.1	97.6±0.2	99.1±0.1	50.44±0.82	65.32±0.70
Reptile	97.68±0.04	99.48±0.06	89.43±0.14	97.12±0.32	49.97±0.32	65.99±0.58
BMAML					53.8±1.46	
VERSA	<b>99.70±0.20</b>	99.75±0.13	<b>97.66±0.29</b>	98.77±0.18	53.40±1.82	67.37±0.86
FBBML(Ours)	99.66±0.21	99.68±0.17	97.49±0.26	98.6±0.19	<b>53.9±1.74</b>	<b>68.7±0.93</b>

匹配网络(Matching Nets)<sup>[23]</sup>由特征提取网络、长短时记忆网络、距离度量网络和注意力分类网络组成。模型无关元学习(MAML)<sup>[24]</sup>是一种双循环的基于梯度优化的方法,其每次外循环对应机器学习训练阶段的一轮训练过程,内循环找到每个任务的梯度,外循环对所有任务梯度进行平均后更新网络参数。原型网络(Prototypical Nets)<sup>[25]</sup>通过一个深层神经网络将  $D$  维的样本数据映射到  $M$  维的嵌入空间,然后在该空间内计算每类样例的均值(原型点),通过比较原型点与查询集样本距离来进行分类。mAP-SSVM 和 mAP-DLM 两种方法<sup>[26]</sup>都是从信息检索(Information Retrieval)的角度来解决小样本学习问题,与本文方法分成支持集和查询集的方式不同,mAP-SSVM 和 mAP-DLM 将 batch 内的每个样本都看作一个查询样本,并且从其他所有样本中检索出与当前查询样本同一类别的样本。网络输出的是其他样本与当前查询样本相似度的排序结果,为评价输出结果的优劣,采用平均准确率(mAP)作为优化的目标函数。Meta-SGD<sup>[27]</sup>方法在 MAML 基础上,除了训练一个较好的初始化参数,还训练了参数的更新方向以及更新步长的大小。SNAIL<sup>[28]</sup>是将时序卷积和注意力机制结合应用到元学习的一种模型方法。BMAML<sup>[29]</sup>是基于概率的 MAML 元学习框架,从概率的角度扩充了 MAML 方法。VERSA<sup>[30]</sup>是一种利用支持集和摊销网络构建分类器参数分布,从参数分布中采样分类器参数对查询集进行分类的方法。

由表 1 可以看到,本文方法在 mini-ImageNet 数据集上得到了最佳的性能,在 Omniglot 数据集上的实验对比中,本文方法没有取得最佳性能,原因主要在于该数据集上实验

方法的整体准确率已经很高,本文方法采用结构中有采样的模块,可以提高模型的泛化能力,但也给模型引入了一定随机性,因此没有取得最佳的性能表现,但 Omniglot 数据集上的准确率依然能充分说明本文方法和纤维丛框架的有效性。Omniglot 数据集的实验充分说明,当实验数据信息充足时本文方法具有一定的随机性,能够避免 FBBML 模型陷入局部最优解。FBBML 模型具有更强的表示能力,能够避免小样本的过拟合问题。

图 6 和图 7 分别为 mini-ImageNet 数据集上 5-way 1-shot 和 5-way 5-shot 实验的准确率与迭代次数的折线图。因训练结果数据过多,因此选每 200 次迭代后的结果进行展示。其中上方红色曲线为训练集准确率,下方蓝色曲线为验证集准确率。由图 6 和图 7 可知,FBBML 在 5-way 1-shot 实验上在 20000 次迭代时已经趋向收敛;在 5-way 5-shot 实验上在 40000 次迭代后趋向收敛;同时可以看出,FBBML 模型并不存在过拟合问题。

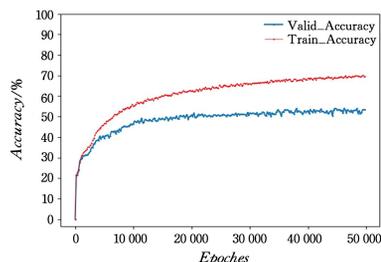


图 6 5-way 1-shot 准确率迭代折线图(电子版为彩色)

Fig. 6 5-way 1-shot accuracy with epochs

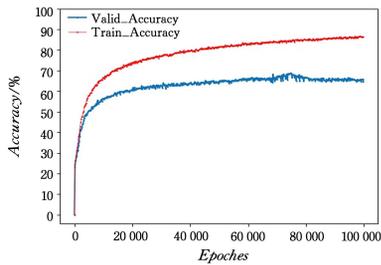


图7 5-way 5-shot 准确率迭代折线图(电子版为彩色)

Fig. 7 5-way 5-shot accuracy with epochs

图8和图9分别为mini-ImageNet数据集上5-way 1-shot和5-way 5-shot实验的准确率与迭代次数的折线图,同样取200次迭代结果的平均损失。

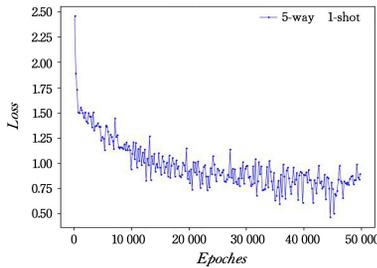


图8 5-way 1-shot 损失迭代折线图

Fig. 8 5-way 1-shot loss with epochs

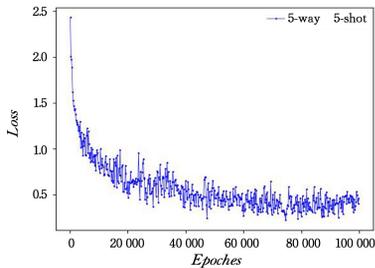


图9 5-way 5-shot 损失迭代折线图

Fig. 9 5-way 5-shot loss with epochs

本文FBBML模型主要针对的是小样本学习,因为样本数量少,所以小样本学习存在容易过拟合的问题。由图8和图9可以看出,FBBML模型在5-way 1-shot的前20000次迭代模型损失快速下降;在5-way 5-shot的前60000次迭代模型损失快速下降。

如图8所示,FBBML模型在45000次迭代后损失值达到最小。同时,图6所示FBBML模型在45000次迭代后模型验证集准确率不再增加。因此,可以确定FBBML模型在5-way 1-shot实验中在45000次迭代后模型收敛,取得最优。如图9所示,FBBML模型在78000次迭代后损失值达到最小,后续训练模型损失值轻微上升。同时,如图7所示FBBML模型在78000次迭代时验证集准确率达到最高,随后模型验证集准确率基本不再上升。由此可以确定,本模型在5-way 1-shot实验中在45000次迭代后模型收敛,取得最优。

对于模型损失震荡问题,本模型训练时在5-way 1-shot实验中学习率为0.00025,在5-way 5-shot实验中学习率为0.0001。学习率本身不是模型损失震荡的原因。本模型损失震荡的主要原因如下:

(1)FBBML模型基于变分贝叶斯理论和纤维丛理论,

通过KL散度(Kullback-Leibler Divergence)<sup>[31]</sup>来计算近似分布 $q$ 与真实分布 $p$ 的相似度,由于KL散度本身的局限性,KL散度本身波动较大。其次当真实分布与近似分布差距过大时,KL散度本身无意义(分母为0),此时KL散度是失效的,考虑到近似分布的初始化为随机值,实验过程中未出现真实分布与近似分布相差过大的情况,但KL散度仍然对本模型具有局限性。因此,可以选择替代的度量函数,例如,考虑使用JS散度(Jensen-Shannon Divergence)<sup>[32]</sup>等度量方法来代替KL散度计算近似分布 $q$ 和真实分布 $p$ 的距离,这是我们后续研究的工作之一。

(2)如第5节所述,本文模型借鉴变分自编码器的思想,引入了重参数技巧来解决分布的采样引起的反向传播难题。因此,重参数层引入了服从(0,1)分布的高斯噪声来增加模型的随机性,这一操作增加了模型的随机性和泛化性能,同样也造成了损失的波动,如何更好地解决采样的反向传播问题同样是我们后续研究的工作之一。

## 6.2 模型复杂度分析

纤维丛元学习框架在N-way k-shot问题下,在元训练阶段需要N类支持集样本,每类样本有k个实例,对于每个task模型需要 $N * k$ 个支持集来完成 $q$ 个Query集的分类,因此模型的每个task的平均复杂度为 $O(N * k * q)$ ;在元测试阶段只需对 $q$ 个查询集样本进行分类,所以每个task的平均复杂度为 $O(q)$ 。

FBBML模型在mini-ImageNet数据集上的实际运行时间如表2所列。

表2 mini-ImageNet数据集上模型的运行时间

Table 2 Time on mini-ImageNet

Test	Time/s
5-way 1-shot	13320
5-way 5-shot	27122.4

FBBML模型在Omniglot数据集上的实际运行时间如表3所列。

表3 Omniglot数据集上模型的运行时间

Table 3 Time on Omniglot

Test	Time/s
5-way 1-shot	6408
5-way 5-shot	9216
20-way 1-shot	8460
20-way 5-shot	30132

由于在训练过程中元学习首先采样出task,因此模型的训练时间与上文推导的时间复杂度具有正向关性,同时模型的训练时间也验证了在本文框架下模型复杂度确实随着分类类别和学习次数(shot)的变化而变化。

**结束语** 本文利用纤维丛理论,从微分几何和流形学习的角度对元学习的机理进行了新的解释。并将纤维丛和流形的概念引入元学习。在此基础之上,本文给出了纤维丛元学习的通用框架,提高了元学习的可解释性。

在接下来的工作中,我们将考虑如何把切丛引入元学习,同时考虑到特征提取维度对流形学习的影响,将选择更佳的特征提取器以及更佳的本真维度来设计具有更好性能的元学习方法。

## 参 考 文 献

- [1] HOSPEDALES T, ANTONIOU A, MICAELLI P, et al. Meta-learning in neural networks; a survey[J]. arXiv:2004.05439, 2020.
- [2] SAHA S, GAN Z, CHENG L, et al. Hierarchical Deep Learning Neural Network (HiDeNN): An artificial intelligence (AI) framework for computational science and engineering[J]. Computer Methods in Applied Mechanics and Engineering, 2021, 373:113452.
- [3] TORREY L, SHAVLIK J. Transfer learning[M]// Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI global, 2010:242-264.
- [4] GLASS G V. Primary, secondary, and meta-analysis of research [J]. Educational Researcher, 1976, 5(10):3-8.
- [5] MAUDSLEY D B. A theory of meta-learning and principles of facilitation: an organismic perspective[D]. University of Toronto, 1980.
- [6] CHAN P K, STOLFO S J. Scaling learning by meta-learning over disjoint and partially replicated data[C]// Ninth Florida AI Research Symposium. 1996.
- [7] BENSUSAN H, GIRAUD-CARRIER C G, KENNEDY C J. A higher-order approach to meta-learning [C]// Inductive Logic Programming, International Conference. London, UK, 2000.
- [8] VILALTA R, DRISSI Y. A perspective View and Survey of Meta-Learning[J]. Artificial Intelligence Review, 2002, 18:77-95.
- [9] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks [C]// Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR.org, 2017.
- [10] SEUNG H S. The manifold ways of perception [J]. Science, 2000, 290(5500):2268-2269.
- [11] SILVA V D, TENENBAUM J B. Global versus local methods in nonlinear dimensionality reduction[C]// Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS). Cambridge, MA, USA: MIT Press, 2002:721-728.
- [12] KÜHNEL W. Differential geometry[M]. American Mathematical Soc., 2015.
- [13] ZHANG P, BAI Y, WANG D, et al. Few-shot Classification of Aerial Scene Images via Meta-learning [J]. Remote Sensing, 2021, 13(1):108.
- [14] ZHANG J, LI F Z. Research on fiber bundle model based on Manifold Learning[J]. Journal of Nanjing University: Natural Science Edition, 2008, 44(5):477-485.
- [15] LU M, LI F. Survey on lie group machine learning[J]. Big Data Mining and Analytics, 2020, 3(4):235-258.
- [16] LI F, ZHANG L, ZHANG Z. Lie group machine learning[M]. Walter de Gruyter GmbH & Co KG, 2018.
- [17] KINGMA D P, WELING M. Auto-encoding variational bayes [J]. arXiv:1312.6114, 2013.
- [18] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4):541-551.
- [19] ZHANG Y, YANG Q. A survey on multi-task learning[J]. arXiv:1707.08114, 2017.
- [20] LU J S. Application of Monte Carlo method in integral solution [J]. Mathematical Learning and Research: Teaching and Research Edition, 2017(5):39.
- [21] LAKE B M, SALAKHUTDINOV R, GROSS J, et al. One shot learning of simple visual concepts[C]// Proceedings of the Annual Meeting of the Cognitive Science Society. 2011.
- [22] RAVI S, LAROCHELLE H. Optimization as a model for few-shot learning[C]// Proceedings of the International Conference on Learning Representations (ICLR). 2017.
- [23] VINYS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning[C]// Advances in Neural Information Processing Systems. 2016:3630-3638.
- [24] FINN C, XU K, LEVINE S. Probabilistic model-agnostic meta-learning[C]// Advances in Neural Information Processing Systems. 2018:9516-9527.
- [25] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]// Advances in Neural Information Processing Systems. 2017:4080-4090.
- [26] TRIANTAFILLOU E, ZEMEL R, URTASUN R. Few-shot learning through an information retrieval lens[C]// Advances in Neural Information Processing Systems. 2017:2255-2265.
- [27] LI Z, ZHOU F, CHEN F, et al. Meta-sgd: Learning to learn quickly for few shot learning[J]. arXiv:1707.09835, 2017.
- [28] MISHRA N, ROHANINEJAD M, CHEN X, et al. A Simple Neural Attentive Meta-Learner[C]// International Conference on Learning Representations. 2018.
- [29] YOON J, KIM T, DIA O, et al. Bayesian model-agnostic meta-learning[C]// Conference and Workshop on Neural Information Processing Systems. 2018.
- [30] GORDON J, BRONSKILL J, BAUER M, et al. Meta-learning probabilistic inference for prediction[C]// International Conference on Learning Representations. 2019.
- [31] ZHANG Y J. Image Information Fusion [M]// Handbook of Image Engineering. Springer, Singapore, 2021:1493-1512.
- [32] NIELSEN F. On a Generalization of the Jensen-Shannon Divergence and the Jensen-Shannon Centroid [J]. Entropy, 2020, 22(2):221.



**LIU Yang**, born in 1996, postgraduate. His main research interests include meta-learning and so on.



**LI Fan-zhang**, born in 1964, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include Lie group machine learning, and dynamic fuzzy logic.