EasyChair Preprint

№ 10372

Meta-Learning with Differentiable Convex Optimization

Yue Gong

June 10, 2023

北京航空航天大學

BEIHANG UNIVERSITY

计算机科学方法论小论文

可微凸优化的元学习

Meta-Learning with Differentiable Convex Optimization

2023 年 6 月

可微凸优化的元学习

摘 要

许多用于小样本学习的元学习方法依赖于简单的基础学习器，例如最近邻分类器。然而，即使在少样本的情况下，经过区分训练的线性预测器也可以提供更好的泛化。本文使用这些预测器作为基础学习器来学习小样本学习的表示，并证明了它们在一系列小样本识别基准中在特征大小和性能之间提供更好的权衡。本文的目标是学习在新类别的线性分类规则下可以很好地泛化的特征嵌入。为了有效地解决目标，本论文利用了线性分类器的两个特性，即凸问题最优条件的隐式微分和优化问题的对偶公式，提出了名为 MetaOptNet 的元学习方法，使得能够在适度增加计算开销的情况下使用具有改进泛化的高维嵌入，在 miniImageNet、tieredImageNet、CIFAR-FS 和 FC100 等小样本学习基准数据集上的实验结果表明了该方法的有效性。

关键词：元学习，小样本学习，凸优化

目 录

1 问题与背景

1.1 研究背景

人类能够很简单的从几个有限的样例中提取事物的特征并且进行区分，但是这对于现代机器学习来说还是很大的挑战。经典的元学习模型用来解决小样本学习问题，包括两个部分，分别是将输入域映射到特征空间的嵌入模型，和将特征空间映射到目标变量的基本学习器。

虽然目前有很多可以选择的基础学习器，但最临近分类器以及其变体是最常用的（例如[1-2]）。然而线性分类器能很好的利用更加丰富的反面数据，更好的学习类别的边界，从而使其表现优于最邻近分类器。

本文研究了以线性分类器作为基础学习器的元学习问题。由于元学习的目标是让模型具有良好泛化性，因此需要在不同任务之间最小化泛化误差，这通常需要使用循环优化的方法来训练线性分类器，带来了很大的计算量。因此可计算性是此问题的关键。

然而，线性模型的目标函数是通常是凸的，因此这个问题可以被有效解决。在小样本的环境下，凸优化可以使元学习变得高效。本文观察到凸的性质中引出的两个额外特性，优化的隐式可微性和分类器的低秩特性[3-4]。第一个特性允许使用一个已有的凸优化模型估计最优值，并隐式地微分最优性条件或卡罗需-库恩-塔克（Karush-Kuhn-Tucker，KKT）条件来训练嵌入模型。第二个特性是对于小样本学习，对偶形式中的待优化变量数目远小于特征维数，通过构造对偶优化问题可以大大减少优化变量的个数。

1.2 国内外研究

元学习探究学习器在不同任务上的泛化能力的影响因素[5-7]。用于少样本学习的元学习方法可以大致分为三组一是基于梯度的方法[8-9]，它通过梯度下降方法寻找和修改嵌入模型的参数。二是最近邻方法[1-2]，它在样本的嵌入特征上学习基于距离的预测规则。三是基于模型的方法[10-11]，学习一个参数化的预测器来估计模型参数。

本文的工作主要与用后向传播进行过程优化的技术相关。Domke[12] 提出了一种基于固定步数的梯度下降和自动微分计算梯度的通用方法。但是由于需要计算梯度，优化器的优化过程中间值需要被记录，这会需要很大的存储空间，应用于规模较大的问题是不现实的。然而，优化器的轨迹（中间值）需要存储以计算梯度，这可能会对大问题造

成限制。Maclaurin 等人考虑了存储开销问题[13]，他们研究了深度网络优化轨迹的低精度表示。如果可以在分析上找到优化的最小值，例如无约束的二次最小化问题，分析的计算梯度也可以被接受。这个成果已经应用于低层视觉问题中[14-15]。

本文的方法使用线性分类器，因为它能够规划为凸学习问题。特别是对于目标函数是一个二次规划问题（QP），可以基于梯度技术高效的获得全局最优解。此外，凸问题的解可以由它们的 KKT 条件所描述，这使得我们可以使用隐函数定理通过学习者[16] 进行反向传播。具体而言，本文使用了 Amos 和 Kolter 的公式化方法[17]，该方法提供了计算 QP 及其梯度的高效 GPU 程序。虽然他们将这个框架应用于学习约束满足问题的表示，但由于出现的问题规模通常很小，因此它也非常适合少样本学习。

2 解决方案

2.1 问题描述

给定训练集 $D^{train}=\{(x_t,y_t)\}_{t=1}^{T}$，基学习器 A 的目标是利用参数 θ 对预测器 $y=f(x,\theta)$ 进行估计, 以在未见的测试集 $D^{test}=\{(x_a,y_a)\}_{t=1}^{Q}$ 上实现更好的泛化能力，

$$\theta=A(D^{train};\phi)=\arg\min_{\theta}L^{base}(D^{train};\theta,\phi)+R(\theta) \tag{2.1}$$

其中，L^{base} 是损失函数，$R(\theta)$ 是正则项，在训练数据有限的情况下，正则项在模型的泛化方面扮演很重要的角色。

为了最小化泛化误差，少样本元学习方法旨在学习任务分布中的最优模型，这可以看作是在一个任务集合上进行学习：$T=\{(D_i^{train},D_i^{test})\}_{i=1}^{I}$，通常被称为元训练集。本文的目标是学习一个嵌入模型 ϕ，使得在给定基础学习者 A 的情况下，在不同任务中达到最小化泛化（或测试）误差的效果。

为了实现这一目标，本文的学习目标是：

$$\min_{\phi}\mathbb{E}_T[L^{meta}(D^{test};\theta,\phi),\text{where}\,\theta=A(D^{train},\phi)] \tag{2.2}$$

图2.1 展示了单一任务的训练和测试过程。一旦学习到嵌入模型 f_ϕ，它的泛化性能可以在一个保留的任务集合（通常称为元测试集）上进行评估。元测试集 $S=\{(D_j^{train},D_j^{test})\}_{j=1}^{J}$

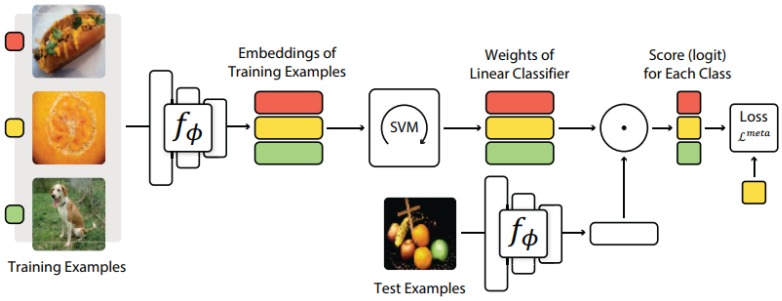

图 2.1 Overview of our approach.

可以用公式来计算：

$$\mathbb{E}_S[L^{meta}(D^{test};\theta,\phi),\text{where}\,\theta=A(D^{train};\phi)] \tag{2.3}$$

根据之前的研究[8-9]，公式 2.2 和公式2.3 中期望值的估计分别被称为元训练和元测试阶段。

2.2 具体方法

2.2.1 任务集

少样本学习使用 K-way，N-shot 分类任务对模型进行评估，其中 K 表示类别数，N 表示每个类别的训练样本数，这被称为任务集。一个任务集 $\tau_i=(D_i^{train},D_i^{test})$ 可以按以下方式进行采样[2, 8]：

总体类集为 C^{train}, 对于每个任务集，首先类 C_i(包含来自 C^{train} 的 K 类) 被有放回抽样得到，训练集 $D_i^{train}=\{(x_n,y_n)|n=1,...,N\times K,y_n\in C_i\}$(每个类包含 N 个图像) 被采样，测试集 $D_i^{test}=\{(x_n,y_n)|n=1,...,Q\times K,y_n\in C_i\}$（每个类包含 Q 个图像）被采样。

需要满足 $D_i^{train}\cap D_i^{test}=\emptyset$，以优化泛化误差。以同样的方式从 C^{val} 和 C^{test} 各自构造元验证集（meta-validation）和元测试集（meta-test）。为了度量嵌入模型对未见类的泛化，$C^{train},C^{val},C^{test}$ 需被互斥选择。

2.2.2 凸基学习器

本文考虑基于多类线性分类器的基学习器（例如支持向量机（SVM）、逻辑回归和岭回归）[18-19]，其中基学习器的目标是凸的。例如，K 类线性 SVM 可以写成 $\theta=\{w_k\}_{k=1}^K$

的形式。Crammer 和 Singer[18] 提出的多类支持向量机的公式是：

$$\begin{aligned} &\theta = A(D^{train};\phi) = \arg\min_{\{w_k\}}\min_{\{\xi_k\}} \frac{1}{2}\Sigma_k\|w_k\|_2^2 + C\Sigma_n\xi_n \\ &\text{subject to} \\ &w_{y_n}\cdot f_\phi(x_n) - w_k\cdot f_\phi(x_n) \geq 1 - \delta_{y_n,k} - \xi_n, \forall n,k \end{aligned} \tag{2.4}$$

这个公式中的 $D^{train} = \{(x_n, y_n)\}$ 表示训练集，其中每个样本 x_n 都有一个真实标签 y_n。C 是一个正则化参数，用于控制模型的复杂度和泛化能力。, 是克罗内克（Kronecker） 函数.

SVM 目标函数的梯度

从图2.1 中可以看出，为了实现端到端的训练，本文需要对 SVM 求解器的解进行微分，以便计算出 $\{\frac{\partial\theta}{\partial f_\phi(x_n)}\}_{n=1}^{N\times K}$。由于 SVM 的目标是凸优化问题，因此具有唯一的最优解，可以在最优（KKT）条件下使用隐函数定理来获得所需的梯度。本文还推导了该凸优化问题的隐函数定理形式[20]，考虑以下凸优化问题：

$$\begin{aligned} &\text{minimize } f_0(\theta, z) \\ &\text{subject to} f(\theta, z) \leq 0 \\ &h(\theta, z) = 0 \end{aligned} \tag{2.5}$$

其中向量 $\theta \in \mathbb{R}^d$ 是问题的优化变量，向量 $z \in \mathbb{R}^e$ 是优化问题的输入参数，即在本文的情况下是 $\{f_\phi(x_n)\}$。本文可以通过求解以下拉格朗日函数的鞍点 $(\tilde{\theta}, \tilde{\lambda}, \tilde{\nu})$ 来优化目标：

$$L(\theta, \lambda, \nu, z) = f_0(\theta, z) + \lambda^T f(\theta, z) + \nu^T h(\theta, z) \tag{2.6}$$

换言之，本文可以通过解决 $g(\tilde{\theta}, \tilde{\lambda}, \tilde{\nu}, z) = 0$ 来获得目标函数的最优解，其中

$$g(\theta, \lambda, \nu, z) = \begin{bmatrix} \nabla_\theta L(\theta, \lambda, \nu, z) \\ \mathbf{diag}(\lambda) f(\theta, z) \\ h(\theta, z) \end{bmatrix} \tag{2.7}$$

对于一个函数 $f(x) : \mathbb{R}^n \to \mathbb{R}^m$，将 $D_x f(x)$ 表示为它的 Jacobian 矩阵 $\in \mathbb{R}_{m\times n}$.

定理 1

（来自 Barratt[3]）假设 $g(\tilde{\theta}, \tilde{\lambda}, \tilde{\nu}, z) = 0$ 。当所有导数都存在时，

$$D_z\tilde{\theta} = -D_\theta g(\tilde{\theta}, \tilde{\lambda}, \tilde{\nu}, z)^{-1} D_z g(\tilde{\theta}, \tilde{\lambda}, \tilde{\nu}, z) \tag{2.8}$$

通过应用隐函数定理于 KKT 条件，本文得到了最优解 $\tilde{\theta}$ 对输入数据梯度的闭合形式表达式，这是凸问题相对于通用优化问题的优势之一。由此，本文不需要反向传播整个优化轨迹来计算梯度，也不需要消耗过多的内存。由于最优解的唯一性，这种方法是可行的。

对偶学习

由于公式2.4中的目标对偶规划本身具有处理嵌入维度上的低依赖性，因此可以重写为如下形式：令

$$w_k(\alpha^k) = \sum_n \alpha_n^k f_\phi(x_n) \quad \forall k. \tag{2.9}$$

可以在对偶空间优化

$$\begin{aligned} &\max_{\alpha^k}[-\frac{1}{2}\sum_k \| \omega_k(\alpha^k) \|_2^2 + \sum_n \alpha_n^{y_n}] \\ &\text{subject to} \quad \alpha_n^{y_n} \le C,\ \alpha_n^k \le 0 \quad \forall k \ne y_n, \\ &\sum_k \alpha_n^k = 0 \quad \forall n. \end{aligned} \tag{2.10}$$

本文可以将公式2.4重写为一个在对偶变量 $\{\alpha^k\}_{K=1}^K$ 上的二次规划（QP），为了解决对偶二次规划，本文使用了一个可微的基于 GPU 的 QP 求解器[17]。

同时，Bertinetto [20] 也采用了岭回归作为基学习器，对于岭回归，优化问题也是一个 QP，因此可以在本文的框架中实现：

$$\max_{\alpha^k}[-\frac{1}{2}\sum_k \| \omega_k(\alpha^k) \|_2^2 - \frac{\lambda}{2}\sum_k \| \alpha^k \|_2^2 + \sum_n \alpha_n^{y_n}] \tag{2.11}$$

其中 w_k 定义同公式2.9。

2.2.3 元学习目标

为了评估模型的性能，本文可以将公式2.2的元学习目标重新表达为：

$$L^{meta}(D^{test};\theta,\phi,\gamma)=\sum_{(x,y)\in D^{test}}[-\gamma\omega_y\cdot f_\phi(x)+\log\sum_k\exp(\gamma\omega_k\cdot f_\phi(x))] \tag{2.12}$$

其中 $\theta=A(D^{train};\phi)=\{\omega_k\}_{k=1}^K$，$\gamma$ 是一个可学习的缩放参数。

3 实验与分析

3.1 实验设置

在元学习设置方面，本文在实验中使用了一个 ResNet-12 网络，遵循[10, 21]。用于 ImageNet 衍生数据集的网络架构为：R64-MP-DB(0.9,1)-R160-MP-DB(0.9,1)-R320-MP-DB(0.9,5)-R640-MP-DB(0.9,5)，而用于 CIFAR 衍生数据集的网络架构为：R64-MP-DB(0.9,1)-R160-MP-DB(0.9,1)-R320-MP-DB(0.9,2)-R640-MP-DB(0.9,2)。本文使用带有 0.9 的 Nesterov 动量和 0.0005 的权重衰减的 SGD 作为优化器。学习率最初设置为 0.1，然后在第 20、40 和 50 个 epoch 时分别更改为 0.006、0.0012 和 0.00024，这是遵循[22] 的实践。

在元训练期间，本文采用了水平翻转、随机裁剪和颜色（亮度、对比度和饱和度）扰动数据增强，如[22-23] 中所述。本文在两个阶段都使用 5 路分类，这是遵循最近的工作[21-22]。对于原型网络，本文将元训练 shot 设置为与元测试 shot 相匹配，这是遵循[1, 22] 的做法。对于 SVM 和岭回归，本文观察到保持元训练 shot 高于元测试 shot 可以获得更好的测试准确性，如图3.1 所示。因此，在元训练期间，本文针对 ResNet-12 的 miniImageNet 将训练 shot 设置为 15；对于使用 4 层 CNN 的 miniImageNet（在表3.3中）将训练 shot 设置为 5；对于 tieredImageNet，将训练 shot 设置为 10；对于 CIFAR-FS，将训练 shot 设置为 5；对于 FC100，将训练 shot 设置为 15。

3.1.1 在 ImageNet 的衍生数据集上的实验

miniImageNet 数据集[2] 是用于少样本图像分类的标准基准测试，由 ILSVRC-2012 [24] 中随机选择的 100 个类组成。tieredImageNet 基准测试[25] 是 ILSVRC-2012 [24] 的一个较大子集，由 608 个类别组成，分为 34 个高级类别。表3.1总结了 5-way miniImageNet 和

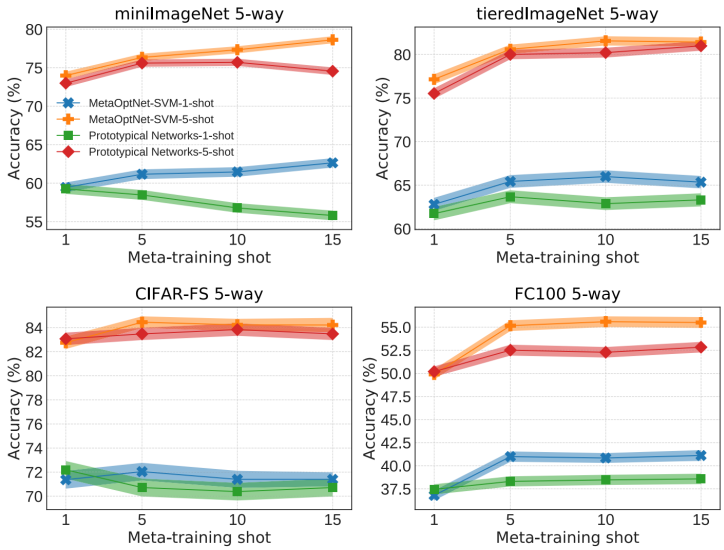

图 3.1 使用不同的元训练样本数量，在 miniImageNet 元测试集上的测试准确率（%）。

tieredImageNet 上的结果。在 miniImageNet 和 tieredImageNet 元测试集上，使用 95% 置信区间的平均 few-shot 分类准确率（%）。“a-b-c-d”表示每个层中具有 a，b，c 和 d 个滤波器的 4 层卷积网络。† 表示使用元训练集和元验证集的并集来元训练元学习器。“RR”代表岭回归。本文的方法在 5-way miniImageNet 和 tieredImageNet 基准测试上取得了最先进的性能。

3.1.2 CIFAR 派生数据集的实验

CIFAR-FS 数据集[20] 是最近提出的 few-shot 图像分类基准测试，由 CIFAR-100 [26] 中的全部 100 个类别组成。FC100 数据集[21] 是另一个源自 CIFAR-100 [26] 的数据集，包含 100 个类别，这些类别被分成 20 个超类。表3.2总结了 5-way 分类任务的结果，本文的 MetaOptNet-SVM 方法实现了最先进的性能。

3.2 实验结果比较

表3.3显示了本文改变两种不同嵌入架构的基学习器后的结果。当本文使用标准的 4 层卷积网络，特征维度较低（1600）时，最近邻类均值分类器[27] 在低维特征下表现良好，如 Prototypical Networks[28] 所示。然而，当嵌入维度远高于 16000 时，SVM 比其他基学习器产生更好的 few-shot 准确性。因此，当高维特征可用时，正则化线性分类器提供了鲁棒性。

对于 ResNet-12，与最近类平均分类器相比，岭回归基学习器的额外开销约为 13%，

表 3.1 与之前在 **miniImageNet** 和 **tieredImageNet** 上的工作进行比较

		miniImageNet 5-way		**tieredImageNet 5-way**	
model	**backbone**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
Meta-Learning LSTM*	64-64-64-64	43.44 ± 0.77	60.60 ± 0.71	-	-
Matching Networks*	64-64-64-64	43.56 ± 0.84	55.31 ± 0.73	-	-
MAML	32-32-32-32	48.70 ± 1.84	63.11 ± 0.92	51.67 ± 1.81	70.30 ± 1.75
Prototypical Networks*	64-64-64-64	49.42 ± 0.78	68.20 ± 0.66	53.31 ± 0.89	72.69 ± 0.74
Relation Networks*	64-96-128-256	50.44 ± 0.82	65.32 ± 0.70	54.48 ± 0.93	71.32 ± 0.78
R2D2	96-192-384-512	51.2 ± 0.6	68.8 ± 0.1	-	-
Transductive Prop Nets	64-64-64-64	55.51 ± 0.86	69.86 ± 0.65	59.91 ± 0.94	73.30 ± 0.75
SNAIL	ResNet-12	55.71 ± 0.99	68.88 ± 0.92	-	-
Dynamic Few-shot	64-64-128-128	56.20 ± 0.86	73.00 ± 0.64	-	-
AdaResNet	ResNet-12	56.88 ± 0.62	71.94 ± 0.57	-	-
TADAM	ResNet-12	58.50 ± 0.30	76.70 ± 0.30	- -	
Activation to Parameter†	WRN-28-10	59.60 ± 0.41	73.74 ± 0.19	-	-
LEO	WRN-28-10	61.76 ± 0.08	77.59 ± 0.12	66.33 ± 0.05	81.44 ± 0.09
MetaOptNet-RR (ours)	ResNet-12	61.41 ± 0.61	77.88 ± 0.46	**65.36 ± 0.71**	**81.34 ± 0.52**
MetaOptNet-SVM (ours)	ResNet-12	62.64 ± 0.61	78.63 ± 0.46	**65.99 ± 0.72**	**81.56 ± 0.53**
MetaOptNet-SVM-trainval (ours)†	ResNet-12	**64.09 ± 0.62**	**80.00 ± 0.45**	**65.81 ± 0.74**	**81.75 ± 0.53**

表 3.2 在不同的元训练样本量下，**CIFAR-FS** 和 **FC100** 元测试集的测试准确率（以百分比表示）

		CIFAR-FS 5-way		**FC100 5-way**	
model	**backbone**	**1-shot**	**5-shot**	**1-shot**	**5-shot**
MAML*	32-32-32-32	58.9 ± 1.9	71.5 ± 1.0	-	-
Prototypical Networks*†	64-64-64-64	55.5 ± 0.7	72.0 ± 0.6	35.3 ± 0.6	48.6 ± 0.6
Relation Networks*	64-96-128-256	55.0 ± 1.0	69.3 ± 0.8	-	-
R2D2	96-192-384-512	65.3 ± 0.2	79.4 ± 0.1	-	-
TADAM	ResNet-12	-	-	40.1 ± 0.4	56.1 ± 0.4
ProtoNets (our backbone)	ResNet-12	**72.2 ± 0.7**	83.5 ± 0.5	37.5 ± 0.6	52.5 ± 0.6
MetaOptNet-RR (ours)	ResNet-12	**72.6 ± 0.7**	**84.3 ± 0.5**	40.5 ± 0.6	55.3 ± 0.6
MetaOptNet-SVM (ours)	ResNet-12	**72.0 ± 0.7**	**84.2 ± 0.5**	41.1 ± 0.6	55.5 ± 0.6
MetaOptNet-SVM-trainval (ours)	ResNet-12	**72.8 ± 0.7**	**85.0 ± 0.5**	**47.2 ± 0.6**	**62.5 ± 0.6**

表 3.3 基础学习器和嵌入网络架构的影响

	miniImageNet 5-way				tieredImageNet 5-way			
	1-shot		5-shot		1-shot		5-shot	
model	acc. (%)	time (ms)	acc. (%)	ime (ms)	acc. (%)	ime (ms)	acc. (%)	ime (ms)
4-layer conv (feature dimension=1600)								
Prototypical Networks	53.47±0.63	6±0.01	70.68±0.49	7±0.02	54.28±0.67	6±0.03	71.42±0.61	7±0.02
MetaOptNet-RR (ours)	53.23±0.59	20±0.03	69.51±0.48	27±0.05	54.63±0.67	21±0.05	72.11±0.59	28±0.06
MetaOptNet-SVM (ours)	52.87±0.57	28±0.02	68.76±0.48	37±0.05	54.71±0.67	28±0.07	71.79±0.59	38±0.08
ResNet-12 (feature dimension=16000)								
Prototypical Networks	59.25±0.64	60±17	75.60±0.48	66±17	61.74±0.77	61±17	80.00±0.55	66±18
MetaOptNet-RR (ours)	61.41±0.61	68±17	**77.88±0.46**	75±17	**65.36±0.71**	69±17	**81.34±0.52**	77±17
MetaOptNet-SVM (ours)	**62.64±0.61**	78±17	**78.63±0.46**	89±17	**65.99±0.72**	78±17	**81.56±0.53**	90±17

SVM 基学习器的额外开销约为 30-50%。

3.3 减少元过拟合

为了缓解过拟合，类似于[23, 29]，本文使用元训练集和元验证集的并集来元训练嵌入，保持超参数（例如 epoch 数）与先前设置相同。表3.1和表3.2显示了使用增强的元训练集（称为 MetaOptNet-SVM-trainval）的结果。本文的结果表明，使用更多的元训练“类”进行元学习嵌入有助于减少对元训练集的过拟合。

表3.4显示了正则化方法对 MetaOptNet-SVM 与 ResNet-12 的影响。本文发现如果不使用正则化，则 ResNet-12 的性能会降低到表3.3中每层 64 个过滤器的 4 层卷积网络的性能水平。这表明正则化对于元学习非常重要。

表 3.4 消融研究

Data Aug.	Weight Decay	Drop Block	Label Smt.	Larger Data	1-shot	5-shot
					51.13	70.88
√					55.80	75.76
	√				56.65	73.72
√	√				60.33	76.61
√	√	√			61.11	77.40
√	√	√	√		62.64	78.63
√	√	√	√	√	64.09	80.00

3.4 双重优化效率

为了验证双重优化确实是有效和高效的，本文在 QP 求解器的不同迭代次数下测量了元测试集上的准确性。结果显示在图 3.2 中。QP 求解器在只进行一次迭代的情况下就达到了岭回归目标的最优解。如[20]。此外，本文观察到对于 1-shot 任务，QP SVM 求解器在 1 次迭代中就达到了最优准确率。这些实验表明，在少样本学习的情况下，解决 SVM 和岭回归的对偶目标非常有效。

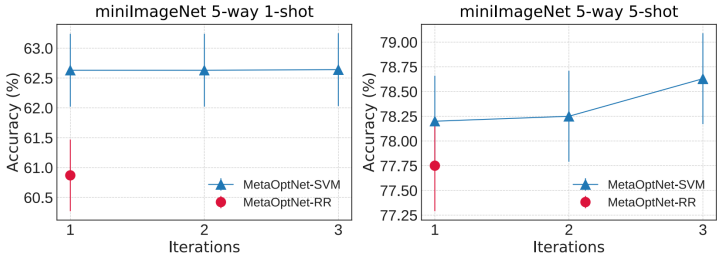

图 3.2 使用不同的元训练样本数量，在 miniImageNet 元测试集上的测试准确率（%）。

4 结论与展望

4.1 结论

本文实现了一种基于凸基学习器的元学习方法，用于少样本学习。通过利用对偶形式和 KKT 条件，可以实现计算和内存高效的元学习，特别适用于少样本学习问题。与最近邻分类器相比，线性分类器在适度增加计算成本的情况下提供更好的泛化能力（如表3.3所示）。本文的实验表明，正则化线性模型可以在减少过拟合的同时实现显著更高的嵌入维度。

4.2 未来展望

本文提出的将正则化线性分类器作为基础学习器在泛化性上具有较强优势。未来的研究方向是探索其他凸基学习器作为基础学习器在元学习模型中的应用，例如核 SVM 等，探究更进一步优化时间复杂度的方法。这将允许更多的训练数据可用于任务，从而逐步增加模型容量的能力。

参 考 文 献

[1] SNELL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[J]. Advances in neural information processing systems, 2017, 30.

[2] VINYALS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning[J]. Advances in neural information processing systems, 2016, 29.

[3] BARRATT S. On the differentiability of the solution to convex optimization problems[J]. arXiv preprint arXiv:1804.05098, 2018.

[4] GOULD S, FERNANDO B, CHERIAN A, et al. On differentiating parameterized argmin and argmax problems with application to bi-level optimization[J]. arXiv preprint arXiv:1607.05447, 2016.

[5] SCHMIDHUBER J. Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook[D]. : Technische Universität München, 1987.

[6] THRUN S. Lifelong learning algorithms.[J]. Learning to learn, 1998, 8:181-209.

[7] VILALTA R, DRISSI Y. A perspective view and survey of meta-learning[J]. Artificial intelligence review, 2002, 18:77-95.

[8] RAVI S, LAROCHELLE H. Optimization as a model for few-shot learning[C]// International conference on learning representations. 2017.

[9] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International conference on machine learning. : PMLR, 2017: 1126-1135.

[10] MISHRA N, ROHANINEJAD M, CHEN X, et al. A simple neural attentive meta-learner [J]. arXiv preprint arXiv:1707.03141, 2017.

[11] MUNKHDALAI T, YUAN X, MEHRI S, et al. Rapid adaptation with conditionally shifted neurons[C]//International conference on machine learning. : PMLR, 2018: 3664-3673.

[12] DOMKE J. Generic methods for optimization-based modeling[C]//Artificial Intelligence and Statistics. : PMLR, 2012: 318-326.

[13] MACLAURIN D, DUVENAUD D, ADAMS R. Gradient-based hyperparameter optimization through reversible learning[C]//International conference on machine learning. :

PMLR, 2015: 2113-2122.

[14] TAPPEN M F, LIU C, ADELSON E H, et al. Learning gaussian conditional random fields for low-level vision[C]//2007 IEEE Conference on Computer Vision and Pattern Recognition. : IEEE, 2007: 1-8.

[15] SCHMIDT U, ROTH S. Shrinkage fields for effective image restoration[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 2774-2781.

[16] KRANTZ S G, PARKS H R. The implicit function theorem: history, theory, and applications[M]. : Springer Science & Business Media, 2002.

[17] AMOS B, KOLTER J Z. Optnet: Differentiable optimization as a layer in neural networks [C]//International Conference on Machine Learning. : PMLR, 2017: 136-145.

[18] CRAMMER K, SINGER Y. On the algorithmic implementation of multiclass kernel-based vector machines[J]. Journal of machine learning research, 2001, 2(Dec):265-292.

[19] WESTON J, WATKINS C, et al. Support vector machines for multi-class pattern recognition.[C]//Esann: volume 99. 1999: 219-224.

[20] BERTINETTO L, HENRIQUES J F, TORR P H, et al. Meta-learning with differentiable closed-form solvers[J]. arXiv preprint arXiv:1805.08136, 2018.

[21] ORESHKIN B, RODRÍGUEZ LÓPEZ P, LACOSTE A. Tadam: Task dependent adaptive metric for improved few-shot learning[J]. Advances in neural information processing systems, 2018, 31.

[22] GIDARIS S, KOMODAKIS N. Dynamic few-shot visual learning without forgetting[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4367-4375.

[23] QIAO S, LIU C, SHEN W, et al. Few-shot image recognition by predicting parameters from activations[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7229-7238.

[24] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115:211-252.

[25] REN M, TRIANTAFILLOU E, RAVI S, et al. Meta-learning for semi-supervised few-shot classification[J]. arXiv preprint arXiv:1803.00676, 2018.

[26] KRIZHEVSKY A, NAIR V, HINTON G. Cifar-10 (canadian institute for advanced re-

search)[Z]. 2010.

[27] MENSINK T, VERBEEK J, PERRONNIN F, et al. Distance-based image classification: Generalizing to new classes at near-zero cost[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(11):2624-2637.

[28] SUNG F, YANG Y, ZHANG L, et al. Learning to compare: Relation network for few-shot learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1199-1208.

[29] RUSU A A, RAO D, SYGNOWSKI J, et al. Meta-learning with latent embedding optimization[J]. arXiv preprint arXiv:1807.05960, 2018.