**Collect and Select Semantic Alignment Metric Learning for Few-Shot Learning**

会议：ICCV2019

代码：<https://github.com/haofusheng/SAML>

**摘要**

小样本学习旨在从很少的训练示例中学习潜在模式，并在实践中显示出希望。但是，在现有方法中直接计算查询图像和支持图像之间的距离可能会引起歧义，因为主要对象可以位于图像上的任何位置。 为了解决这个问题，本文提出了一种针对小样本学习的语义对齐度量学习（SAML）方法，该方法通过“收集并选择”策略对齐语义相关的主导对象。 具体来说，我们首先计算一个关系矩阵（RM）以“收集”从查询图像中提取的3D张量的每个局部区域对和支持图像的平均张量的距离。 然后，注意技术适用于“选择”语义相关的对，并对其施加更多的权重。 此后，利用多层感知器（MLP）将重加权的RM映射到其相应的相似度评分。 理论分析证明了SAML的泛化能力，并提供了理论保证。 实证结果表明，实现了语义对齐。 在基准数据集上进行的大量实验验证了该方法的优势，并证明了SAML明显优于当前的最新方法。

**1.简介**

与通常需要大规模数据集的常规方法（例如ImageNet）它解决了这样一个问题：收集如此大量的数据非常耗时，在实践中有时是不现实的[37]。最近，通过神经网络从图像中提取的特征已经表现出在许多计算机视觉任务中的深刻表现能力[42，20，23，14]。 基于从图像中提取的3D张量，度量学习方法极大地促进了小样本学习的前沿。 具体而言，度量学习首先计算分别从查询图像和支持图像中提取的3D张量之间的距离，然后基于该距离学习分类器。

然而，大多数现有方法[48，37，16] 按元素索引直接计算每个张量对之间的距离度量。这种直接比较可能会带来严重的歧义，因为主要对象可以位于图像上的任何位置。 因此，可能会将一幅图像中的主要对象与另一幅图像中在语义上无关的局部区域进行比较（请参见图1中红色和蓝色框的比较）。

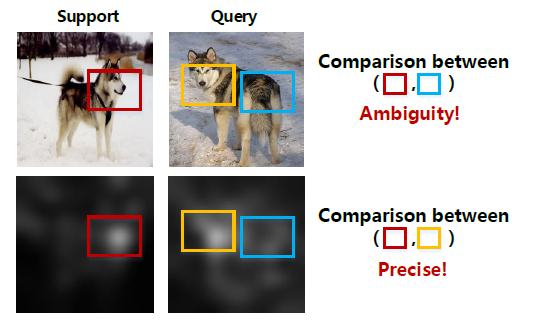


图1.显示的两个图像属于同一类别，而关键对象（狗）出现在不同的位置。 根据空间索引直接计算两幅图像之间的距离会引入歧义，即在狗的头部（红色框）和狗的尾巴（蓝色框）之间成对出现。 所提出的方法SAML将局部区域与相同的语义信息对齐（请参见红色和黄色方框之间的比较）。

为了解决这个问题，本文提出了一种语义对齐度量学习（SAML）方法，通过“收集和选择”策略对齐图像上语义相关的局部区域。 具体来说，SAML首先从关系矩阵（RM）中“收集”所有本地区域对与查询图像和支持图像的距离。 每个局部区域由卷积神经网络（CNN）从相应图像提取的3D张量中的向量表示。 然后，SAML使用语义技术“选择”语义相关的局部区域对，并根据相关性重新称重它们。 然后，由多层感知器（MLP）处理相关权重的RM，以计算相似性分数，以确定查询图像是否来自支持类。

理论分析评估了SAML的泛化能力，并提供了O（1 /√N）泛化范围，而没有明显依赖于嵌入网络和MLP的参数大小（N是训练样本集中的事件数）。 在两个标准基准数据集CUB [40]和miniImageNet [34，31]上进行的大量实验表明，SAML明显优于最新方法。

**2.相关作品**

现有的小样本学习作品主要来自以下四个类别：度量学习，元学习，幻觉和基于注意力的学习。

**度量学习**：基于度量学习的方法具有相同的范例：（1）通过嵌入网络将所有图像（包括支持和查询图像）映射到表示空间，并计算每个支持类别的表示； （2）计算每个查询图像到所有支持类的距离； （3）将每个查询实例分配给与其最接近的支持类。

现有的方法主要集中在前两个步骤中的一个，因为第三步是相对完善的：（1）嵌入网络的设计已从早期的孪生神经网络[19]发展为有条件移位的神经元的快速适应[27]。 和内存匹配网络[5]。 最近，为了更好地捕获几何信息，[48，37]引入了3D张量。 类表示的计算可以追溯到[16]，其使用为每个支持类别嵌入深层特征的平均值作为其表示。 最近，乔等人。 目前通过预训练的神经网络的激活来预测每种新型支持类的表示[33]； （2）进行距离分类的距离度量的设计源于余弦距离[31]和欧几里德距离[16]到更高级的距离测量，例如由Graph Neural Networks [39]计算的距离测量。

对于大多数现有作品而言，一个重要的问题是，距离是通过根据元素索引直接计算得出的[48，37]，并且经常引入严重的歧义，将语义上不相关的部分进行比较。

**元学习**：基于元学习的方法可以学习学习算法本身。 Ravi等。 提出了一种基于LSTM的元学习器，该学习器学习精确的优化算法，以便在少数情况下训练神经网络分类器[34]。 Finn等。 设计MAML来训练一个元学习器，该元学习器提供良好的参数初始化，这样只有很少的更新就可以快速学习新颖的任务[8]。 Meta-SGD调整更新方向和学习率以快速适应新任务[22]。 不过，这些方法通常需要昂贵的高阶梯度，当扩展到更深的网络体系结构时，这可能会导致失败如[26]中所示。 相应地，Mishra等。 结合时间卷积和软注意力，提出了可扩展到更深层网络架构的通用金属学习器架构[26]。 另外，一阶优化[29]和潜在嵌入优化[36]技术已经被提出来解决这个问题。

**幻觉**：基于幻觉的方法通过两个不同的方向来增加每个新类别中带标签的实例的数量，从而解决了小样本学习问题。Bharath等人的早期工作。 应用类别无关的转换为每个新颖类别生成尽可能多的新实例[13]。 其后续发展[43]利用了元学习学习的最新进展，该进展共同优化了元学习者和幻觉者，从而产生了高质量的训练实例。 张等提出MetaGAN生成与从特定任务[47]的真实数据无法区分的样本，和Akshay等人将生成的假实例视为强大的正则化器[24]。

**基于注意的**：最近，注意机制已被引入到小样本学习中。 例如，Yan等。 [44]利用空间注意力来定位相关对象区域和任务注意力以选择相似的训练数据进行标签预测，从而提出了基于两种注意力机制的双重注意力网络。任等人。 [35]提议通过注意力吸引者网络来规范新类的学习。 Hu等[15]提出了一种基于注意力的多上下文指导（A-MCG）网络，该网络在支持和查询分支之间集成了多尺度上下文特征，从支持集中加强了更好的指导。 相比之下，我们采用注意力机制来“选择”语义上相关的区域。

**3.语义对齐度量学习**

本节介绍了我们提出的方法SAML。 在描述图像嵌入之前，我们首先回顾问题定义以进行小样本学习。 然后，我们描述语义收集的“收集和选择”。最后，提供了度量的两个实例。

**3.1问题定义**

在这项工作中，我们集中于M-way K-shot问题，其中M是类别数，K是每个类别中的示例数（K通常是一个小整数，例如1或5）。

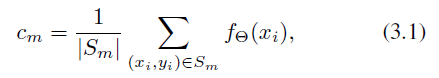
小样本学习数据集由三部分组成：训练集，验证集和测试集，它们的标签空间彼此不相交（例如，在验证或测试期间看不到训练期间看到的类别）。 通常，每个集合都包含丰富的类别和示例，它们远大于M和K。自[31]提出以来，这三个集合通常分为许多情节，每个情节都包含一个支持集S = 和查询集。 支持集和查询集都是从训练/验证/测试集中随机抽取的。 此外，S和Q是不相交的（S = Q =），而它们共享相同的标签空间。

为了模拟现实世界中的短镜头学习场景，所有的训练，验证和测试程序都在情节上实施。 例如，在每个训练迭代中，从训练集中随机采样一个情节训练参数更新。 此过程重复多次，直到模型收敛到稳定状态。 剧集的验证和测试类似。

**3.2 图像嵌入**

通过神经网络对每个图像xi进行卷积，我们可以获得一个3D张量来表示该图像，其中f是神经网络学习的假设函数，为神经网络的参数，是由所有3D张量形成的表示空间，C，H和W分别是张量的三个维度的长度。 通过这种方式，我们将所有图像嵌入到表示空间中。 每个3D张量中都有H×W个C维像元，每个像元都是对应图像中区域的局部区域特征（LRF）（也是接收场）。 与1D [16，31]或具有其他尺寸的特征相比，3D张量可以更好地捕获几何信息，因此是基于度量学习的小样本学习方法的常见选择。 图像嵌入可以通过许多神经网络来实现。有关我们的嵌入网络的详细信息，请参阅第5.1节。

一集中每个支持类的K张图像。 当K> 1时，一项重要的任务是从K张单个图像的3D张量计算支撑类的表示形式。 在本文中，我们利用K 3D张量的经验平均值来表示相应的支持类别：



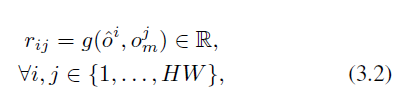
其中cm是第m个支持类的类表示，Sm是情节中的类m的支持集，而| Sm | 是Sm中的示例数。 支持类表示cm也位于表示空间。 与单个图像的表示相似，每个类表示cm也由H×W C维特征作为LRF构成。

为了便于解释，将表示cm的H×W LRF表示为。 类似地，每个查询图像的3D表示的LRF为。

**3.3收集并选择语义对齐**

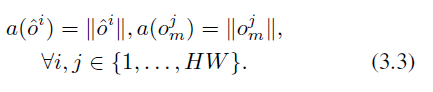
图像中确定其类别的本地区域可以位于任何地方。 例如，对于狗图像，包含狗的行列式局部区域可能位于一个图像的右上角，而在另一个图像的中心区域。 因此，直接根据位置索引计算它们之间的距离可能会在语义上不相关的局部区域配对，并可能导致严重的歧义[48，37]（见图1）。 因此，在语义上对齐表示是计算距离的重要任务。

为了解决这个问题，我们提出了一种“收集和选择”策略来实现语义对齐。 具体来说，我们计算由查询图像中的一个LRF和支撑类中的一个LRF组成的所有LRF对的度量距离。 所有距离值都收集在关系矩阵（RM）中，并根据空间索引进行定位。 例如，RM *R*的（i，j）元素rij是由查询图像〜xn的第i个LRF 与支持类表示形式的第j个LRF之间的某个度量g定义的距离 ：



其中rij是一个标量，反映了两个LRF的相似性。度量可以根据各种情况而变化。 有关度量实例的讨论在第3.4节中提供。 RM携带查询图像和支持类别之间相似度的所有信息。 获取RM的过程称为该策略的“收集”阶段。

RM包含语义上不相关的局部区域的距离。 注意技术增强了关键对象，同时抑制了背景[46，4，17]。 通过使用注意技术，我们可以更加注意语义相关的LRF对。 在本文中，我们选择了基于激活的注意力[46]，其中每个LRF的范数都被定义为注意力值a：



我们重新加权距离rij

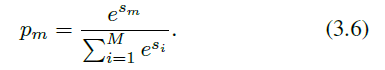


这样，可以抑制语义上不相关的局部区域的距离； 同时，增加了语义相关局部区域的距离。 因此，我们实现了语义对齐。 采用注意力技术来找到语义上相关的局部区域的过程称为该策略的“选择”阶段。 之后，将重新加权的RM R'馈送到MLP以计算相似性评分（因子）以执行进一步分类：

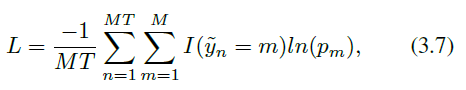


其中表示MLP的可学习参数。

对于每个查询实例，有M个相似度得分sm分别表示查询实例与所有支持类的相似性。 链接所有 M个相似度得分构成一个向量，我们形成判别函数。 为了执行最终分类，我们使用softmax函数来计算分配给第m类的测试示例的概率pm：



根据概率，我们进一步定义损失函数：

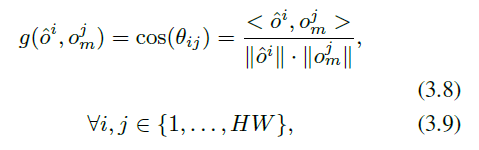


其中I（·）是指示符函数，如果其参数为true，则等于1；否则为0；〜yn是〜xn的标号，T是每个查询集中的实例数。

**3.4指标的实例化**

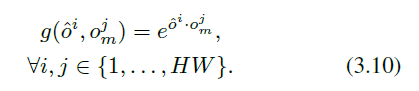
上一小节中的度量功能有很多选择。 本文使用两个简单的度量标准进行实验：余弦度量标准和高斯度量标准。 实验表明，简单的指标足以令人满意。

余弦度量：余弦距离定义为两个特征之间的角度的余弦：



其中是ˆoi和oj m之间的夹角，<·，·>是内积，是范数。 它对于面部验证[41]和图像分类[12]有效。

高斯度量：高斯函数也可以作为g的选择：



**4.理论分析**

本节从理论角度研究了根据训练样本的大小对我们的方法进行泛化的能力。 我们首先为所提出模型的覆盖范围（覆盖范围）提出一个上限。 覆盖范围控制了由我们提出的方法引起的假设空间的复杂程度。 然后，我们获得了所提出方法的泛化误差（泛化边界）的上限。 泛化界为我们的方法提供了理论上的保证。

小样本学习可以建模为二进制分类问题。 具体来说，每个情节都是一个例子。 查询和支持图像是实例，标签是​​它们是否来自同一类。 在这个部分，我们使用二进制分类的理论来评估我们的方法。 因此，我们提供了理论上的保证，即训练样本集中有多少个情节就足够了。

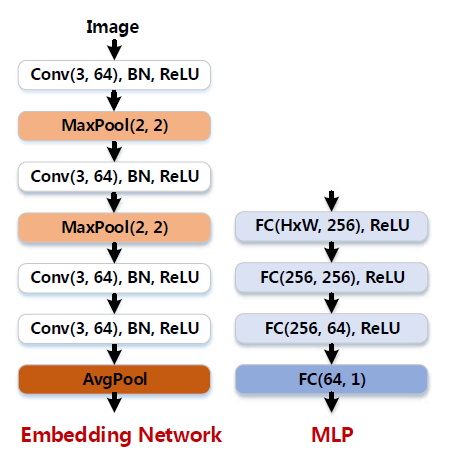


图4.嵌入网络和MLP的网络架构。

如图4所示，我们的模型涉及两个神经网络，即嵌入网络和MLP。 它们通过不影响假设复杂度的固定操作连接。 具体而言，该操作计算了每个支持类的表示形式，并计算了从查询实例和类表示形式得出的关系矩阵。 假设所建议模型的输入为X。嵌入网络由四个卷积层，两个最大池化层和一个平均池化层组成。 我们分别用权重矩阵A1，A2，A3和A4以及非线性和表示它们。 相应地，嵌入网络的输出可以表示为

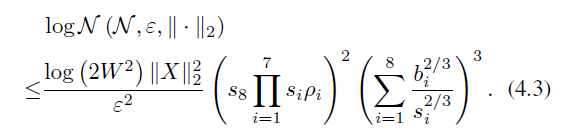


关系矩阵和注意技术的收集是固定的非线性操作。 在这里，我们将其表示为。 此外，我们通过权重矩阵A5，A6，A7和A8以及非线性表示MLP。 因此，我们提出的算法的最终输出为：



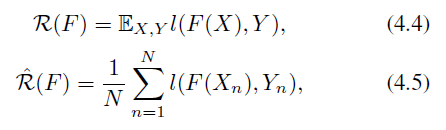
为了简化以下定理，我们定义。 假设我们模型的输出分类器的假设空间为H。则可以得到以下定理。

定理1（覆盖边界）。 假设第i个非线性项的Lipschitz常数为，运算的Lipschitz常数为。 假设每个权重矩阵的光谱范数为：而且，假设每个权重矩阵Ai具有参考矩阵Mi，其满足。 然后，“-覆盖数”满足

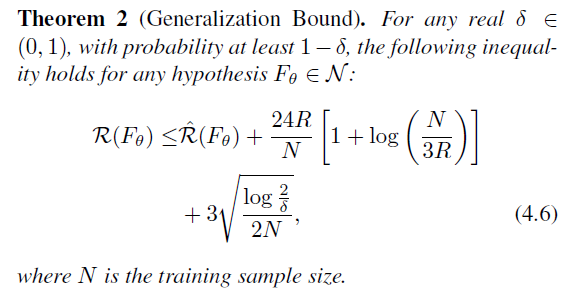


W是整个算法中特征图的最大维度。

在此省略详细的证明，并在附录中给出。根据覆盖范围，我们可以得到以下定理。 为简洁起见，我们表示等式的右侧（RHS）（4.3）为。此外，我们分别将预期风险和经验风险定义为



其中（X，Y）是特征标签对，N是训练样本量，l是损失函数。



通过将定理1应用于学习理论中的两个经典结果，可以直接获得定理2，这里省略了这些定理，但在附录中提供。 附录中还提供了详细的证明。 等式 （4.6）给出了我们提出的算法的推广界。这为我们提出的方法提供了理论上的保证。

**5.实验**

本节介绍实验设置，消融研究以及与最新方法的比较。

**5.1实验设定**

数据集：miniImageNet数据集是ImageNet [7]的子集，由100个类别组成，每个类别包含600个带标签的实例。 我们采用通用分割分别获得用于培训，验证和测试的64、16和20个类别。 CUB数据集最初是为细粒度分类而设计的，由11 788例200多种鸟类组成。 我们将数据集随机分为100个训练，50个验证和50个测试类别。对于miniImageNet和CUB，图像均被调整为224×224，并且不采用任何数据扩充。为了更直观地了解这两个数据集，附录中显示了一些图像。

网络：我们的嵌入网络和MLP的详细信息如图4所示。由于嵌入网络是我们的工作重点，并且为了进行公平的比较，因此我们的嵌入网络与[31，37]具有相似的主干，但仍然有一些小的变化 获得足够的LRF。 例如，仅保留前两个最大池化层，而最后一个最大池化层被平均池化层代替。 miniImageNet和CUB的平均池化层的跨步设置为5，从而得到100个LRF。 另外，相似性分数通常限制在0到1的范围内，这是通过在最后一个完全连接的层之后添加S型函数来实现的。 在这里，我们省略了MLP常用的S型函数。

实现细节：我们以余弦度量为例，介绍SAML的实现细节。 总体流程图如图3所示。在嵌入所有支持图像和查询图像之后，我们合并它们的空间尺寸。 可以通过在重构的LRF上执行矩阵乘积来直接实现SAML，如附录中的图8所示。 所有实验均在PyTorch框架下进行[1]。 我们使用亚当[18]的初始学习率为10−3，每2 000个情节减半。 训练情节的总数为20，000。请注意，在每个测试情节中，每个班级都要测试T = 15个查询图像。

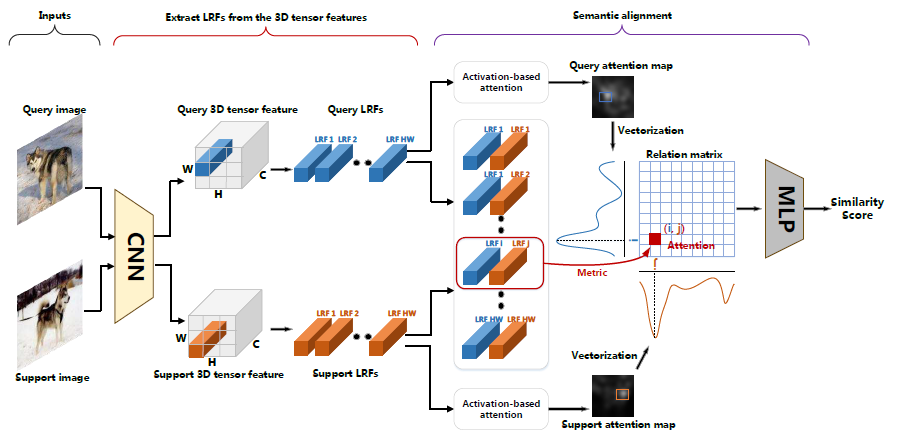


图3.提出的方法SAML的流程图。

**5.2消融研究**

我们研究度量功能，LRF数量，图像大小和注意方法的影响。 另外，讨论了时间复杂度。 实验在mini-ImageNet上进行。

图像尺寸的影响：为了更好地控制LRF的数量，我们采用了224×224而不是84×84的较大图像尺寸。为了公平比较，我们评估了图像尺寸对少拍分类性能的影响，并且 结果如表1所示。我们还使用RelationNet [37]进行了类似的比较。 与RelationNet相比，我们的方法受益于图像大小的增加。图像尺寸越大，LRF重叠越小，并且各个LRF越独立。 直观地，性能随图像尺寸而增加的原因可以归因于感受野与图像尺寸的较小比例。

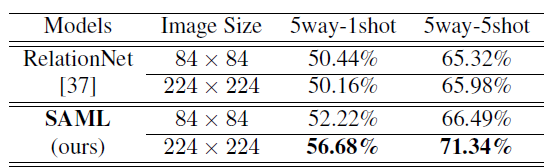


表1.图像大小对小样本分类性能的影响。 实验在mini-ImageNet上进行。

度量函数的影响：如表2所示，我们研究了度量函数的选择对短击学习性能的影响。余弦和高斯无注意的表现要比注意版本差得多，这可以归因于他们对 所有LRF对的比较结果。 对于同一类别，一对相同语义LRF显然比一对语义无关LRF对更重要，因为后者可以增加差异。 注意余弦和注意高斯通过注意方法抑制了语义无关的LRF的比较。 对于不同类别，关键区域显示出显着差异，SAML也可以捕获这些差异。

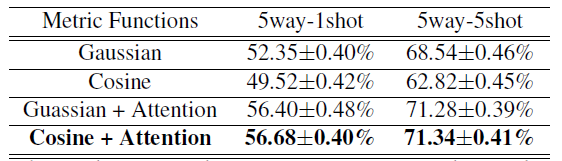


表2.不同度量功能对小样本分类准确性的影响。 实验在mini-ImageNet上进行。

LRF数量的影响：可以通过设置嵌入网络的最终平均池化层的跨度来调整LRF的数量。 我们评估LRF数量对小样本学习性能的影响，结果如图5所示。随着LRF数量的增加，5-way 1次和5-way 5-shot分类的准确性都会提高，并且达到饱和。 当LRF的数量为100时。基于此观察，我们默认在以下实验中将LRF的数量设置为100，并且该设置对mini-ImageNet和CUB都适用。

注意方法的影响：我们研究注意方法的影响。 图2中显示了一些正确分类的图像及其相应的注意图。关键区域用红色矩形标记，所有这些区域对应于相同的语义概念，即dog。 注意图的作用是在增强对象的同时抑制背景。 注意方法和度量函数的组合减少了关键对象与语义无关部分之间的比较所引入的歧义。 但是，当复杂背景中存在不同比例的支配对象时，我们的方法就会受苦，图6中显示了一些故障示例。我们进一步优化了方法来解决此问题。 具体来说，我们在嵌入网络（参见图4）之后引入了一个初始[38]运算符（[2×2最大池，3×3卷积，5×5卷积，7×7卷积]）来提取具有以下特征的对象的特征： 不同的规模。 结果在5.3节中讨论。

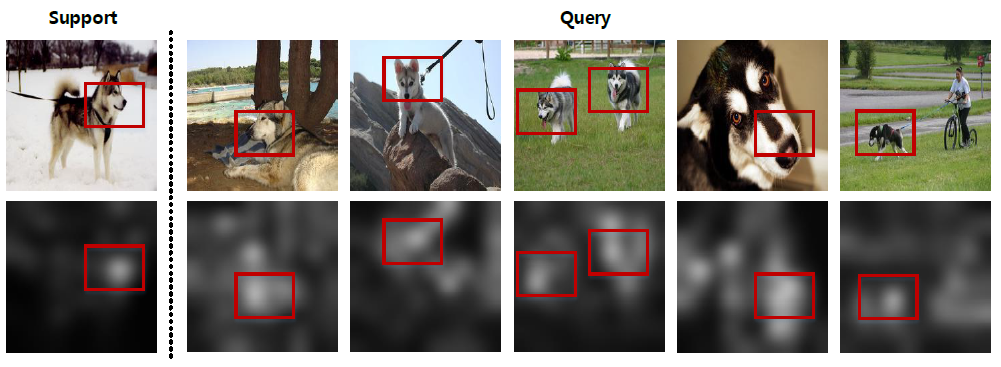


图2.语义相关的局部区域的插图。 他们证明了语义对齐是通过SAML实现的。

时间复杂度：SAML的计算是矩阵乘法（请参见附录中的图8），已在流行的深度学习平台（例如PyTorch [1]和TensorFlow [2]）中进行了优化，并且不再是主要的计算瓶颈 。 我们比较了不同方法的时间成本，结果显示在表3中。我们的方法在时间成本方面具有竞争力（在测试过程中5way-1shot排名第二快，5way-5shot排名第一）。

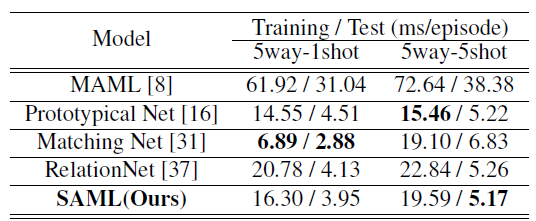


表3. miniImageNet上不同方法的每集培训/测试时间成本。

**5.3与现有技术的比较**

我们将SAML与最先进的方法进行比较。在此，采用余弦度量。 可以在附录中找到更多的经验结果，包括度量标准更复杂的结果。

Mini-ImageNet上的结果：为了进行公平的比较，通过评估方式执行了两个常见任务，即5-way 1-shot和5-way 5-shot分类。我们从mini-ImageNet测试集中随机抽取了600集，然后以95％的置信区间报告了几近分类的准确性。 我们还将重复测试过程10次并报告差异。 结果显示在表4中。对于5-way 1-shot和5-way 5-shot分类，尤其是在5-way 5-shot任务上，我们的方法比最新方法的性能要好得多。 （> 2.4％）。 mini-ImageNet中关键对象的比例差异很大。 当采用尺度不变特征时，针对5-way 1-shot和5-way 5-shot分类的朴素SAML改进分别为1.01％和1.69％。

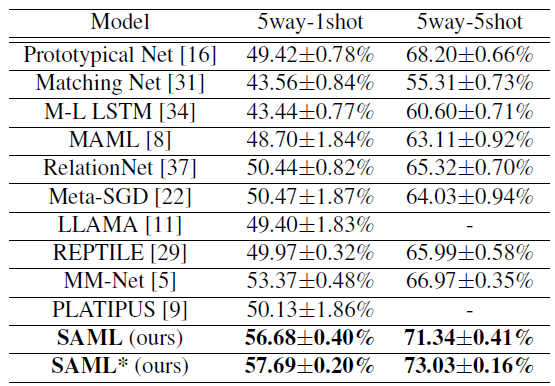


表4. Mini-ImageNet上小样本分类的准确性。 “-”表示“未报告”。 “ \*”表示“采用初始运算符”。

关于CUB的结果：CUB是由不同物种的鸟类组成的细粒度图像分类数据集。 与为通用识别而收集的miniImageNet相比，CUB很简单，因为主要对象始终是鸟类，并且背景相对干净。 但是，这些鸟的位置仍然表现出很大的变异性。 对CUB进行两项任务，即5way-1shot和5way-5shot分类，实验结果如表5所示。我们的方法比现有方法具有更好的性能。具体来说，细粒度的5way1shot分类和5way-5shot分类任务的增量分别为6.88％和2.22％，这是令人惊讶且令人印象深刻的性能提升。 当我们使用给定的CUB边界框裁剪所有图像时，CUB中主要对象的比例大致相同。 因此，采用尺度不变特征对性能几乎没有影响。

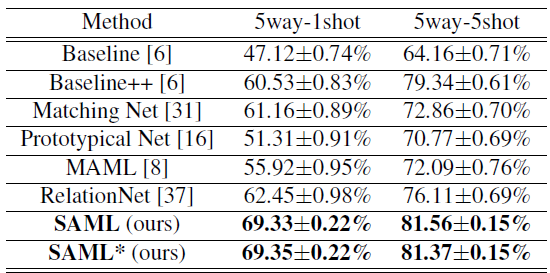


表5. CUB很少分类的准确性。 “ \*”表示“采用初始运算符”。

**6.总结**

主导对象可能出现在图像的任何部分。 因此，直接计算之间的距离根据索引从图像中提取的特征可能会导致严重的歧义，因为我们可能会比较语义上不相关的局部区域。 为此，我们提出了一种语义对齐度量学习（SAML）方法，该方法通过“收集并选择”策略对齐语义相关的本地区域。 具体来说，我们定义一个关系矩阵（RM）以“收集”查询实例的局部区域对和支持类均值的所有距离，然后利用注意力技术“选择”并更加注意语义相关的局部区域对。 实验结果表明，实现了语义对齐。 对泛化范围的理论分析证明，SAML在看不见的数据上的可行性得到了保证。 在标准基准数据集上进行的大量实验通过与最新的少量学习方法进行比较，证明了SAML的优越性。