**Diversity with Cooperation Ensemble Methods for Few-Shot Classification**

会议：ICCV2019

代码：<http://thoth.inrialpes.fr/research/fewshot_ensemble/>

**摘要**

小样本分类包括学习一个预测模型，该模型能够有效地适应新的类别，而仅给出几个带注释的样本。 为了解决这一具有挑战性的问题，元学习已成为一种提倡“学习适应”能力的流行范例。 然而，最近的研究表明，没有元学习的简单学习策略可能具有竞争力。 在本文中，我们进一步走了一步，表明通过解决小样本学习分类器的基本高方差问题，可以大大胜过当前的元学习技术。 我们的方法包括设计一组深度网络以利用分类器的方差，并引入新的策略来鼓励网络合作，同时鼓励预测多样性。在微型ImageNet，tiered-ImageNet和CUB数据集上进行了评估，我们在其中显示了即使通过蒸馏获得的单个网络也可以产生最新的结果。

**1.引言**

卷积神经网络[17]已成为计算机视觉中用于建模图像的标准工具，从而在许多视觉识别任务（例如分类[16]，对象检测[8、19、26]或语义分割[[ 8、20、27]。 大量注释的数据集，例如ImageNet [28]或COCO [18]，似乎在成功中发挥了关键作用。 但是，根据手头的任务，标注大型主体非常昂贵，而且并不总是可行的。 因此，提高深度神经网络的泛化能力并消除对大量注释的需求至关重要。

尽管可以从不同的互补观点（例如大规模无监督学习[4]，自监督学习[7、12]或通过开发专用于深度网络的正则化技术[2、36]）来解决这样一个严峻的挑战 ，我们选择本文重点研究基于集成方法的方差减少原理。

具体来说，我们对小样本分类感兴趣，在这种分类中，首先在中等大小的带注释语料库上从头开始训练分类器，也就是说，无需利用外部数据或预先训练的网络，然后评估其适应新类的能力 ，仅提供很少的带注释的样本（通常为1或5）。 不幸的是，仅在极少样本的情况下对新分类任务进行卷积神经网络的微调已显示出较差的结果[9]，这激励了社区开发专用方法。

小样本学习的主导范式建立在元学习[9，24，30，32，31，33]的基础上，该元学习被制定为学习如何适应新的学习问题的原则。这些方法将大型带注释的语料库划分为分类任务，目标是在各个任务之间传递知识，以提高泛化性。 虽然元学习原则似乎适合小样本学习，但其经验优势尚未明确。 确实有确凿的证据[5、11、23]表明，使用元学习从头训练CNN的效果远比以标准方式训练CNN功能（即通过最小化依赖语料库注释的经典损失函数）进行的效果差。 另一方面，发现学习仅具有元学习功能的最后一层可以产生更好的结果[11，23]。 然后，最近在[5]中表明，基于距离的简单分类器可以达到与元学习方法相似的准确性。

我们的论文进一步走了一步，表明无元学习方法可以得到改进，并且在小样本学习中可以大大超越当前的现有技术水平。我们的攻角包括使用集成方法来减少小样本学习分类器的方差，鉴于注释数量少，这不可避免地会很高。给定一个初始的中等大小的数据集（遵循小样本学习的标准设置），最基本的集成方法包括：首先独立训练几个CNN，然后再冻结它们，并删除最后一个预测层。 然后，给定一个新的类（带有少量带注释的样本），我们为每个网络建立一个平均质心分类器，并根据与质心的距离估算测试样本的类概率（根据基本概率模型）[21，31]。 然后将获得的概率通过网络求平均，从而获得更高的准确性。

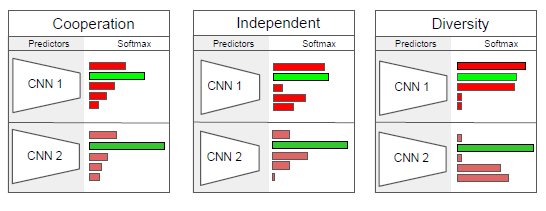


图1：两个网络上的合作和多样性策略的图示。 所有网络都将相同的图像作为输入，并使用softmax计算相应的类概率。 合作鼓励归一化后的非地面真理概率（红色）相似，而多样性则鼓励正交。

虽然我们证明了独立训练网络的基本集成方法已经表现良好，但我们引入了惩罚术语，使网络在训练期间可以进行协作，同时又鼓励了足够的预测多样性，如图1所示。 更轻松的学习和规范化，使集成中的各个网络可以互惠互利。 鼓励多样性的动机是集成方法的经典方法[6]，在这种方法中，一群做出不同预测的弱学习者通常比一个强大的学习者在一起表现更好。 乍一看，这两个原则似乎相互矛盾，但我们证明这两个原则实际上是有用的，并且比基本集成方法所产生的结果要好得多。 最后，我们还表明，通过蒸馏训练的单个网络[14]可以模仿集合的行为，并且效果很好，这在测试时带来了显着的加速。 总而言之，我们的贡献是三方面的：

•我们引入了鼓励合作和多样性以学习网络整体的机制。我们研究这两个原则以进行小样本学习，并描述它们有用的机制。

•我们证明，在不使用元学习的情况下，针对小样本分类的性能可能会大大优于当前的最新技术。

•作为一个较小的贡献，我们还将展示如何通过使用其他未标记的数据将集合提炼到单个网络中，而准确性会略有下降。

**2.相关工作**

在本节中，我们讨论有关小样本学习，元学习和集成方法的相关工作。

小样本分类。 典型的小样本分类问题由称为元训练和元测试的两个部分组成[5]。 在元训练阶段，给了一个足够大的带注释的数据集，该数据集用于训练预测模型。 在元测试期间，提供了新颖的类别以及一些带注释的示例，并且我们评估了预测模型重新训练或适应的能力，然后对这些新类进行了归纳。

元学习方法通​​常会从元训练数据集中抽取小样本学习分类任务，并对模型进行训练，以使该模型可以泛化已搁置的新任务。 例如，在[9]中学习了“良好的网络初始化”，使得针对新问题的少量梯度步骤足以获得良好的解决方案。 在[24]中，作者同时学习了网络初始化和由长期短存储网络（LSTM）表示的更新规则（优化模型）。受深度学习方法流行之前开发的少数​​快照学习策略的启发[21]，还提出了基于到质心距离的基于距离的分类器，例如原型网络[31]，或者更复杂的具有关注度的分类器[33]。 所有这些方法都考虑了经典的骨干网络，并使用元学习从头开始对其进行训练。

最近，发现这些元学习不是最优的[11、22、23]。 具体来说，第一步，通过使用元训练数据在经典分类任务上训练网络，然后在第二步[22、29、35]中仅通过元学习进行微调，从而获得更好的结果。 其他诸如[11，23]的方法只是冻结第一步中获得的网络，并使用元学习训练一个简单的预测层，从而获得相似的性能。 最后，论文[5]演示了基于元距的分类器，无需元学习的简单基线同样可以很好地工作。 我们的论文进一步推动了此类原理的发展，并表明通过适当的减少方差的技术，这些方法可以大大超越当前的技术水平。

组合方法。 众所周知，集成方法可以减少估计量的方差，从而可以提高预测的质量[10]。 为了从平均中获得准确性，通常使用各种随机化或数据增强技术来鼓励高度的预测[3，6]。 尽管整体的单个分类器可能表现不佳，但平均预测的质量有时却令人惊讶地很高。

即使在神经网络训练时使用集成方法成本很高，但事实表明，训练为模仿集成行为的单个网络几乎可以表现良好[14]（一种称为蒸馏的过程），从而消除了测试时的开销。在高度并行的实现情况下提高蒸馏的可扩展性，在[1]中提出了在线蒸馏程序。 在那里，鼓励每个网络都同意该集成的其他网络所做的平均预测，从而获得更稳定的模型。 但是，我们工作的目标截然不同。 他们鼓励在网络之间进行合作的形式确实针对可伸缩性和稳定性（由于工业限制），但是在线提炼网络的性能不一定比基本整体策略更好。 另一方面，我们的目标是提高预测质量，并且比基础集成做得更好。

为此，我们鼓励网络之间的预测以不基于地面事实标签的类别概率进行匹配，从而以不同的方式鼓励合作。 虽然我们表明，仅当网络数量较少时，单独使用这种策略才有用，但是当网络数量增加时，鼓励多样性就变得至关重要。 最后，我们证明蒸馏可以帮助减少测试时的计算开销。

**3.我们的方法**

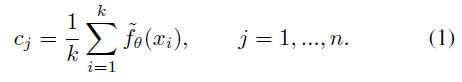
在本节中，我们将从初步组件入手，介绍针对小样本的方法。

**3.1均质中心分类器**

现在，我们解释如何使用固定特征提取器和均质中心分类器执行小样本分类。

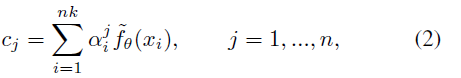
使用原型分类器进行小样本分类。 在元训练阶段，我们将获得带有注释的数据集Db，用于训练由CNN表示的预测函数f。 形式上，在Db上训练CNN之后，我们删除最终的预测层，并将所得的矢量用作给定图像x的一组视觉特征。 参数θ表示网络的权重，该权重在此训练步骤后冻结。

在元测试期间，我们得到一个新的数据集Dq = ，其中n是许多新类别，k是每个类别的可用示例数。 （xi，yi）代表图像标签对。 然后，我们建立一个均值质心分类器，生成类原型

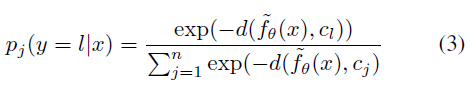


最后，将测试样本x分配给最近的质心。 简单的质心形分类器已被证明在小样本分类的背景下是有效的[5、21、31]，以下实验对此进行了证实。

均质心分类器的动机。 我们在此报告的实验表明，模型比（1）更复杂不一定能为小样本学习带来更好的结果。考虑确实是（1）的参数化版本：



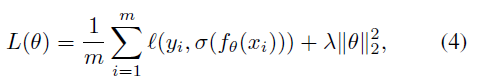
通过最大化概率模型的似然性可以通过梯度下降学习权重αji



其中d（·，·）是距离函数，例如欧几里得距离或负余弦相似度。 由于系数是从数据中获悉的，而不是像（1）中那样任意设置为1 / k，因此，如果适当地进行正规化，可能会期望该方法产生更好的分类器。 当我们对从mini-Imagenet-test采样的1000个5shot学习任务进行上述分类器的评估时（有关该数据集的详细信息，请参见实验部分），我们平均得到相似的结果：（1）vs. 77.28±0.46％。 （2）为77.01±0.50％，这证明在这种非常低的样本方案中学习有意义的参数是困难的。

**3.2学习深度网络集成**

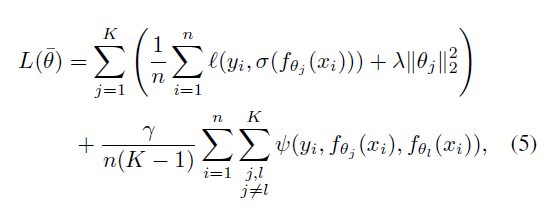
在元训练期间，需要在训练集上最小化以下损失函数：



其中f是和以前一样的CNN。 成本函数ℓ（·，·）是真实标签与预测类别概率p =σ（f（x））之间的互熵，其中σ是归一化指数函数，而λ是权重衰减参数。

当独立训练K个网络的集合时，将分别为每个网络求解（4）。虽然这些术语看起来相同，但是当使用不同的初始化和随机种子进行训练时，深度神经网络提供的解决方案通常会有所不同，这使得集成方法在这种情况下具有吸引力。

在本文中，我们对网络的集成感兴趣，但是我们也想对它的成员之间的关系进行建模。 这可以通过考虑成对惩罚函数ψ来实现，从而得出联合公式：



其中是通过将所有参数θj连接而获得的向量。 通过仔细设计函数ψ并适当设置参数γ，可以实现集成的理想属性，例如预测过程中的预测多样性或协作性。

**3.3鼓励多样性与合作**

为了减少小样本学习分类器的高方差，我们使用经过特殊交互作用函数ψ训练的集成方法，如（5）所示。 然后，一旦在元训练中学习了参数θj，就通过考虑与等式（3）中所示的基本概率模型相关联的K个均质中心分类器的集合来执行元测试中的分类。 给定一个测试图像，对K类概率进行平均。 从经验上看，这种策略比投票方案表现更好。

如我们在实验部分所示，成对关系函数ψ的选择会显着影响集成的质量。 在这里，我们描述了三种不同的策略，它们均以鼓励多样性预测的准则为起点，在不同的制度下均能带来收益。

多样性。 鼓励多样性的一种方法包括在学习过程中引入随机化，例如，通过使用数据增强[3，10]或各种初始化。在这里，我们还评估了直接作用于网络预测的相互作用函数ψ的影响。 给定图像x，用θi和θj参数化的两个模型分别导致类概率和。 在训练过程中，自然会鼓励pi和pj接近中的赋值向量ey，并在位置y处使用单个非零条目，其中y是与x相关的类标签，而d是类数量。

从[14]中我们知道，即使仅使用pi或pj的最大条目进行预测，其他条目（通常不对应于地面真实性标签y）也携带有关网络的重要信息。 然后自然而然地考虑概率取决于不是地面真理标签。 形式上，这些是通过将pi和pj中的y设置为零并将重新归一化相应的向量，使它们求和而获得的。 然后，我们考虑以下分集惩罚



当与损失函数结合使用时，得出的公式会鼓励网络根据真实的标签做出正确的预测，但同时也鼓励它们做出不同的第二，第三，等等选择预测（ 参见图1）。 如实验部分所示，当网络数量很大时，这种惩罚措施特别有效。 通常，它平均会使单个分类器的性能变差，但会使整体预测更加准确。

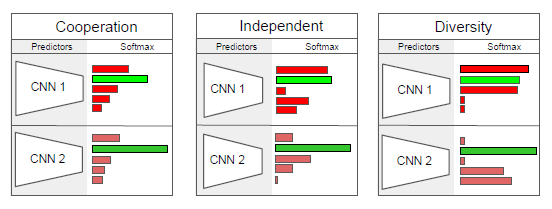


图1：两个网络上的合作和多样性策略的图示。 所有网络都将相同的图像作为输入，并使用softmax计算相应的类概率。 合作鼓励归一化后的非地面真理概率（红色）相似，而多样性则鼓励正交。

合作。 显然与先前的原则相反，鼓励条件概率相似（尽管使用不同的度量标准）也可以通过允许网络进行协作以更好地学习来提高预测质量。 我们的实验表明，单独使用这种原则可能有效，但是当培训网络的数量很少时，它似乎最有用，这表明需要在合作与多样性之间进行权衡。

具体来说，我们的实验表明，使用负余弦（换句话说，与（6）相反）无效。 但是，事实证明，诸如对称的KL-divergence这样的惩罚可提供所需的效果：



通过使用此惩罚，我们设法获得了更稳定，更快速的培训，从而使单个网络的性能更好，而且（也许令人惊讶的）更好的集成。不幸的是，我们还观察到集成的增益随着集成中网络数量的增加而减少，因为单个成员变得太相似了。

稳健与合作。 给定在前两个惩罚下进行的实验，合作与多样性之间的权衡似乎对应于两种机制（网络数量低与网络数量高）。 这促使我们开发一种旨在实现最佳权衡的方法。 在考虑合作惩罚（7）时，我们尝试通过几种其他方式来增加预测的多样性。 i）在每次训练迭代中，我们会从集合中随机删除一些网络，这会使网络在不同的数据流上学习，并降低知识传播的速度。 ii）我们在每个网络中引入了Dropout，以增加随机性。 iii）我们为每个网络提供相同图像的不同（裁剪，彩色）转换，这使集成对于输入图像转换更加鲁棒。 总体而言，发现该策略在大多数情况下效果最佳（请参见图2）。

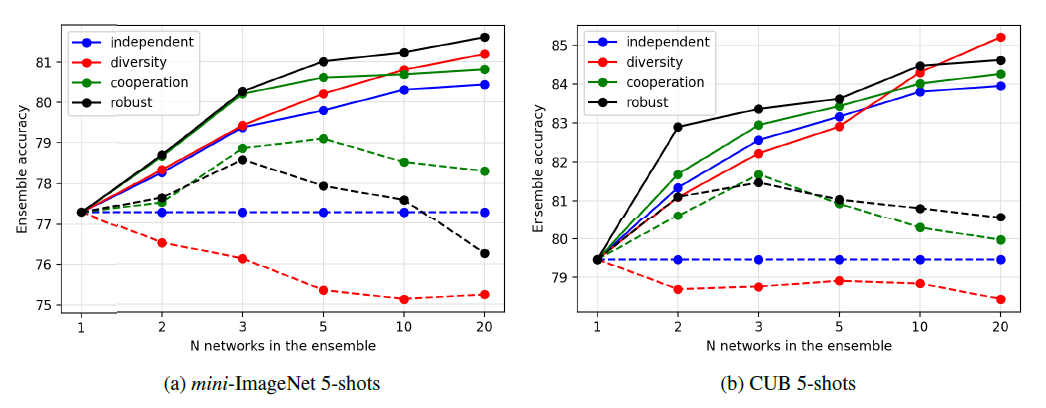
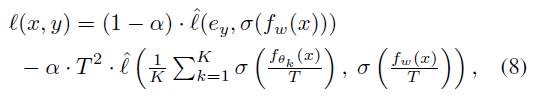


图2：对于各种数量的网络，不同的集成策略（每种颜色一个）的精度。 实线给出了汇总预测后的整体准确性。 集合中单个模型的平均性能用虚线绘制。 彩色效果最佳。

**3.4. Ensemble distillation**

作为大多数集成方法，我们的集成策略在训练时会引入大量的计算开销。 为了消除测试时的开销，我们使用知识蒸馏的变体[14]将集合压缩为单个网络fw。 给定元训练数据集Db，我们在示例（x，y）上考虑以下成本函数：



其中，是交叉熵，ey是真实标签y的单次嵌入。 第二项以参数T进行蒸馏（参见[14]）。 鼓励单模FW与集成的平均输出相似。 在我们的实验中，我们可以获得性能相对接近整体的模型（请参见第4节）

对分布外行为进行建模。 当对数据集Db执行蒸馏时，网络fw会模仿特定数据分布上的集成行为。但是，在测试时会引入新的类别。 因此，我们还尝试使用其他未注释的数据进行蒸馏，从而获得更好的性能。

**4.实验**

我们现在提出实验，以研究合作和多样性对整体方法的影响，并从实验和实施细节入手。

4.1。 实验设置

数据集。 我们使用mini-ImageNet [24]和tiered-ImageNet [25]，它们分别来自原始ImageNet [28]数据集和Caltech-UCSD Birds（CUB）2002011 [34]。 Mini-ImageNet由100个类别组成，其中64个类别用于训练，16个类别用于验证，20个类别用于测试，每个类别有600张图像。 Tiered-ImageNet也是ImageNet的子集，其中包括351个训练类，97个验证类和160个测试类，总共779,165张图像。选择分割，使训练类与测试类完全不同，这与mini-ImageNet不同。 CUB数据集包含200种以上鸟类的11,788张图像。 我们采用[35]中的训练，验证和测试分割，它们最初是通过将所有200个物种随机分为100种进行训练，50种进行验证和50种进行测试而创建的。

评估。 在少拍分类中，测试集用于抽样N个5种分类问题，其中每个类别仅提供k个示例进行训练，提供15个示例进行评估。 我们遵循[9、11、22、23、24]并测试我们的算法，其中k = 1和5，并且N设置为1000。每次，随机抽取类和相应的训练/测试示例。 对于我们所有的实验，我们报告了1000个任务和95％置信区间的平均准确度（以％为单位）。

实施细节。 对于所有实验，我们使用Adam优化器[15]进行初始学习，其初始学习率为10−4，在训练过程中，如果连续p个纪元的验证准确度未见提高，则在训练过程中将其降低10倍。 对于小型ImageNet，我们使用p = 10，CUB数据集使用20。 将集合提炼成一个网络时，p会加倍。 我们在训练过程中使用随机裁剪和增色以及参数λ= 5·10−4的体重减轻。 所有实验均使用ResNet18架构[13]进行，该架构允许我们在单个GPU上训练20个网络的集合。然后，将输入图像重新缩放为224×224的大小，并以16的小批量进行组织。通过对验证集进行5次射击评估来计算验证准确性。 在元测试阶段，我们从图像中提取尺寸为224×224的中心裁剪，并将其输入特征提取器。 测试时不使用其他预处理。 建立平均质心分类器时，将（3）中的距离d计算为负余弦相似度[31]，将其重新缩放10倍。

为了公平比较，我们还评估了由输入图像大小为84×84的ResNet18 [13]和输入图像大小为80×80的WideResNet28 [37]组成的集成。所有详细信息在附录中进行了介绍。

**4.2进行Few-Shot分类的集成**

在本节中，我们将研究采用成对互动方式鼓励合作或多样性的整体训练的效果。 为此，我们分析了微型ImageNet和CUB数据集上的集成大小与其1拍和5拍分类性能之间的联系。

有关这三种策略的详细信息。 联合训练模型时，数据流将在所有网络之间共享，并且权重更新会同时发生。 这是通过将所有模型放在同一GPU上并优化损耗来实现的（5）。 当训练多样化的集成时，我们使用余弦函数（6）并选择参数γ= 1，该参数在对n =5和n = 10个网络的测试值（）中的验证集上表现最佳。 然后，将该值保留为其他n值。

为了加强网络之间的协作，我们使用对称的KL函数（7）并以相同的方式选择参数γ= 10。 最后，用协作关系罚分和相同的参数γ训练鲁棒的集成策略，但是在最后一层之前使用Dropout的概率为0.1。 在每次迭代中，每个网络都以0.2的概率从集合中删除； 不同的网络会收到相同图像的不同变换，即不同的随机裁剪和颜色增强。

结果。 附录的表1和表A1总结了使用我们的策略训练的集成的小样本分类准确度，并与基本集成进行了比较。 在微型ImageNet数据集上，1幅和5幅分类的结果彼此一致。 通过合作训练可以使较小的集成（n≤5）更好地执行，这导致集成成员的个体准确性更高，如图2所示。但是，当n≥10时，与多样性策略相反，合作的效率较低 ，这得益于更大的n。 从图2中可以看出，集成中的各个成员变得更糟，但是集成准确性却大大提高了。

最后，健壮的策略似乎在几乎所有设置中对所有n值都表现最佳。 CUB数据集的情况类似，尽管我们注意到健壮的集成的表现与n = 20时的多样性策略相似。

**4.3蒸馏集成**

我们提取各种大小的健壮集成，以研究随着集成大小增长而产生的知识转移性。 为此，我们使用元训练数据集并使用参数T = 10和α= 0.8优化损失（8）。 对于使用外部数据的策略，我们在每次迭代时将来自COCO [18]数据集的8张图像（不带注释）随机添加到元训练数据的16个带注释的样本中。 这些图像仅对损失的蒸馏部分有所贡献（8）。

附录的表1和表A1分别显示了微型ImageNet和CUB数据集的模型准确性。 对于mini-ImageNet上的5镜头分类，集合与其提炼版本之间的差异非常小（大约1％），而添加额外的未注释数据有助于缩小这一差距。 令人惊讶的是，蒸馏模型的1发分类准确度比其相应的完整集成略高。 在CUB数据集上，即使完整集成的性能持续增长，提炼后的模型在n = 5之后也会停止改善。 这似乎表明可能已达到单个网络的容量，这表明此处使用了更复杂的体系结构。 与这种假设一致，添加额外的数据没有像mini-ImageNet那样有用，这很可能是因为COCO和CUB的数据分布更加不同。

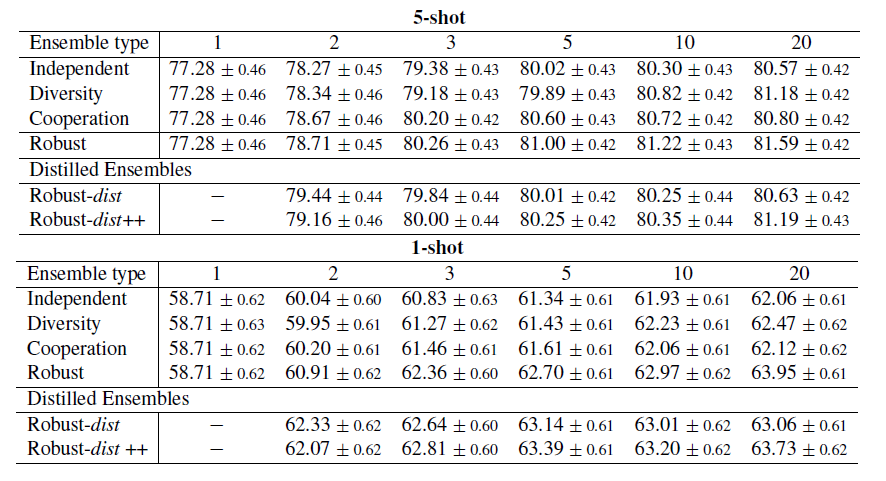


表1：小型ImageNet上的分类准确率很少。 第一列给出了策略，第一行指出了一个集合中网络的数量N。 在这里，dist表示将整体提炼成单个网络，“ ++”表示使用多余的未注释图像进行蒸馏。 我们在mini-ImageNet-test上进行了1000次独立实验，并以95％的置信区间报告了平均值。 所有网络均在mini-ImageNet训练集上训练。

在表2、3中，我们还将蒸馏网络的性能与文献中的其他基准进行了比较，包括当前最新的元学习方法，表明我们的方法在mini-ImageNet上的性能明显更好[24 ]和分层ImageNet [25]数据集。

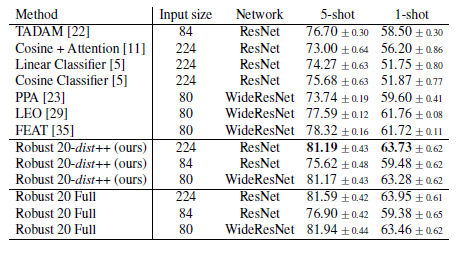


表2：在1次和5次miniImageNet上将蒸馏乐谱与其他方法进行比较。 最后两列显示1拍和5拍学习任务的准确性。 为了评估我们的方法，我们在MiniImageNet-test上进行了1000次独立实验，并报告了平均值和95％的置信区间。 在这里，“ ++”表示使用了额外的未注释图像进行蒸馏。最后一个模型是完整的集合，不应与表的其余部分直接比较。

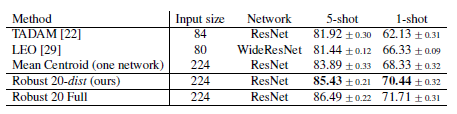


表3：在1-shot和5-shot分层ImageNet上，蒸馏乐谱与其他方法的比较[25]。 为了评估我们的方法，我们在分层ImageNet测试中进行了5000次独立实验，并以95％的置信区间报告了平均准确度。

**4.4关系处罚研究**

有许多方法可以对集合成员之间的关系进行建模。 在本小节中，我们将研究和讨论这样的特定选择。

输入到关系函数。 如[14]所指出的，网络的softmax层获得的分类概率似乎携带了很多信息，并且对蒸馏很有用。但是，在元训练之后，这样的概率通常接近二进制向量，且具有与地面真相标签相关联的主导值。 为了使较小的值更明显，蒸馏使用参数T，如式（8）所示。 给定一个由网络计算的分类概率，我们尝试了一种由引入新概率p =σ（p / T）组成的策略，其中强调了非真实值的贡献。 但是，如果在多样性（6）或合作（7）的惩罚范围内使用，我们发现基本集成方法没有任何改进。 取而代之的是，我们发现，以第3.3节中所述的条件为基础，计算不是基于地面真相标签的类概率会好得多。

这在下面的实验中得到了说明，该实验使用大小为n = 5的两个网络集成。我们在第一个中对全部概率矢量强制执行相似性，并在[1]之后使用T = 10的softmax和有条件的非地面真实性进行计算 第3.3节中定义的第二个概率。 在MiniImageNet上进行测试时，使用合作培训公式时，第二种策略的效果要比第一种提高约1％（79.79％比80.60％）。 使用分集标准已经做出了类似的观察。 相比之下，没有交互的基本集成方法可达到约80％。

关系功能的选择。 原则上，任何相似性措施都可以用来设计惩罚措施鼓励合作。在这里，我们表明，实际上，根据所需的效果（合作或多样性），选择正确的准则来比较概率向量（余弦相似度，L2距离，对称的KL散度）至关重要。 在表4中，当将上述函数插入公式5中并带有特定符号时，我们对MiniImageNet数据集上具有n = 5个网络的集合执行5镜头分类任务的比较。选择每个实验的参数γ，以使验证集的性能最大化。

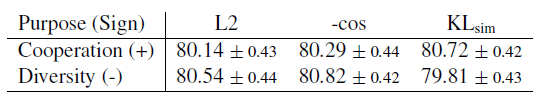


表4：在mini-Imagenet 5镜头上评估不同的关系标准第一行指示使用哪个功能作为关系标准，第一列指示该功能用于哪个目的以及相应的符号。 为了评估我们的方法，我们对CUB测试进行了1000次独立实验，并以95％的置信区间报告了平均准确度。 所有集成都经过mini-ImageNet训练。

寻找多样性时，余弦相似度比负L2距离略好，尽管精度在误差范围内。对各种γ使用负KLsim不能与独立训练区分开，或者对于较大的γ值会损害性能（未在表中报告）。 至于合作，KLsim正比L2距离或负余弦相似性要好。 我们认为，此行为是由于这些函数比较概率矢量中较小值的方式存在重要差异。 负余弦或L2损失会严重损害最大的差异，而KLsim则专注于在一个向量中接近0而在第二个向量中更大的值。

**4.5域转移下的性能**

最后，我们评估了域移位下集成方法的性能。 我们通过对微型ImageNet训练集上的模型进行元训练，并在CUB测试集上评估模型。 [5]首先提出以下设置，其目的是评估当训练和测试分布之间的差异较大时，算法的性能以适应。 为了与原始工作中报告的结果进行比较，我们采用了他们的CUB测试划分。表5将我们的结果与[5]中列出的结果进行了比较。 我们可以看到，完全健壮的集成体及其精简版本都不能比在冻结网络顶部训练线性分类器做得更好。 但是，它比基于距离的方法（表中的余弦分类器表示）要好得多。 但是，如果使用不同的集成，则可以达到最佳精度。 这不足为奇，并突出了在组合弱分类器时具有多种模型的重要性。

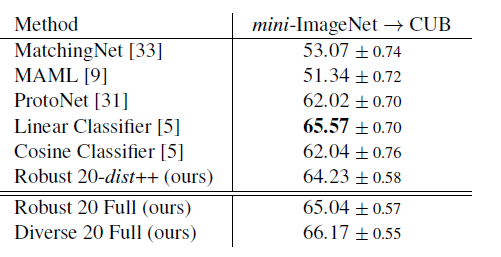


表5：域移位下的5镜头分类精度。 最后两个模型是完整的集合，不应与表的其余部分直接比较。 我们从[5]开始对CUB-test进行了1000次独立实验，并在此处报告了平均值和置信区间。 所有集成都在mini-ImageNet上进行训练。

**5结论**

在本文中，我们显示了用于小样本学习的基于距离的分类器具有高方差，可以通过使用分类器的集合来显着减少。与传统的集成范式不同，在传统的集成范式中，通过各种随机化和数据增强技术来鼓励预测的多样性，我们表明，鼓励网络在训练期间进行协作也很重要。

通过蒸馏获得的单个网络的整体性能（在测试时没有计算开销）导致了小样本学习的最新性能，而无需依赖元学习范式。 虽然这样的结果对于元学习方法来说可能是消极的，但这可能仅意味着在该领域仍有大量工作要做，以真正学习如何学习或适应。