Shaping Visual Representations with Language for Few-Shot Classification

**Abstract**

通过描述世界的特征和抽象，语言是人类学习的重要工具，也是机器学习模型的一个有希望的监督来源。在未充分探索的场景中，我们使用语言来改进少数镜头的视觉分类，在训练期间自然语言任务描述是可用的，但在测试时对于新任务是不可用的。此设置的现有模型在测试时对新描述进行采样，并使用这些模型对图像进行分类。相反，我们提出了语言形状学习（LSL），这是一个端到端的模型，它规范了视觉表现来预测语言。LSL在概念上更简单，数据效率更高，并且在两个具有挑战性的少数镜头域中优于基线。

**1引言**

人类之所以能成为强大而高效的学习者，部分原因在于学习语言的能力（Chopra等人，2019年；Tomasello，1999年）。例如，我们可以了解知更鸟，不是通过看成千上万的例子，而是通过被告知知更鸟是一种红色腹部和棕色羽毛的鸟。这种语言进一步塑造了我们看待世界的方式，限制了我们对新概念的假设：给定一只新鸟（如海鸥），即使没有语言，我们也知道像腹部和羽毛颜色这样的特征是相关的（古德曼，1955）。

在本文中，我们用语言指导视觉表象的学习，研究在测试时没有语言的环境，因为丰富的语言监督往往是无法在野外遇到的新概念。在这种情况下，如何才能最好地使用语言？一种选择是对表示进行正则化训练，以预测语言描述。另一种方法是直接利用语言的合成特性，将其作为离散隐变量模型的瓶颈。例如，最近的潜在语言学习（L3；Andreas等人，2018）模型兼具两种功能：在训练期间，使用语言对图像进行分类；在测试时，在没有语言的情况下，根据语言形状的图像嵌入从解码器中采样描述。

目前还不清楚像L3这样的大多数收益模型的瓶颈还是规范化。我们理清了这些影响，提出了语言形状学习（language shaped learning，LSL），这是一种端到端的模型，它使用语言形状的视觉表示（如图1所示），从而避免了瓶颈。我们发现，离散的瓶颈会影响性能，特别是在有限的语言数据下；相比之下，LSL在架构上更简单、更快，使用语言更有效，并且在两个few-shot传输任务中优于L3和基线。

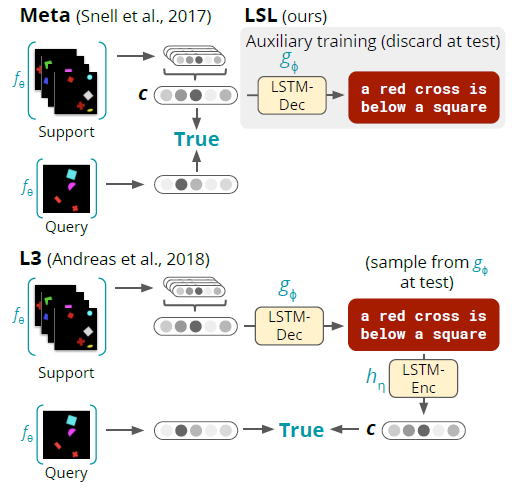


图1：与明确使用语言作为分类瓶颈的模型（Andreas等人，2018）相比，我们提出了few-shot分类模型，这些模型的学习表示被限制在预测训练期间的自然语言任务描述。

**2 Related Work**

语言被证明有助于在各种环境下进行视觉分类，包括不转移的传统视觉分类（He和Peng，2017），以及在测试时以类标签或zero（Frome等人，2013；Socher等人，2013）或few-shot（Xing等人，2019）学习描述的形式提供的语言。与以往的工作不同，我们在测试时没有语言，并且测试任务与培训任务不同，因此培训中的语言不能用作附加的类信息（cf.He和Peng，2017）或对标记额外的域内数据监管较弱（参见Hancock等人，2018）。我们的设置可以看作是使用特权信息学习（LUPI；Vapnik和Vashist，2009）的一个实例，在这里，更丰富的监督只在培训期间增强模型。

在这个框架中，具有属性和其他特定领域的理论基础的学习被广泛地处理（Zaidan等人，2007；Donahue和Grauman，2011；Tokmakov等人，2019）；语言较少所以Gordo和Larlus（2017）使用标题之间的METEOR分数作为图像检索专用嵌入的相似性度量，但不直接使用ground-true解释。斯里瓦斯塔瓦等人。（2017）探索类似于我们的监控设置，除了简单文本和符号域中的描述可以通过语义解析轻松转换为可执行逻辑形式。

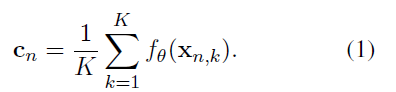
另一个研究方向是研究自然语言对跨语言任务（例如，蕴涵；Camburu等人，2018）和视觉（Hendricks等人，2016，2018）的可解释性的解释，但在这里，我们研究预测语言是否真的能提高任务绩效；在文本中也探索了类似的想法（Rajani等人），以及强化学习（Bahdanau等人，2019年；Goyal等人，2019年）领域。

**3 Language-shaped learning**

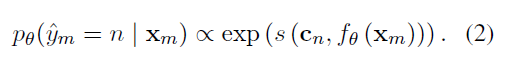
我们感兴趣的是，语言解释可以帮助学习在任务之间更有效地概括的表示，特别是当每个任务的训练数据很少，并且有许多与输入一致的虚假假设时。因此，我们研究了few-shot（元）学习环境，即模型必须从一组数据有限的训练任务中学习，然后将其推广到同一领域中未见的任务。

具体来说，在N-way，K-shot学习中，任务t由N个支持类组成，带有K个例子：。每个任务都有M个查询示例。给定第M个查询示例作为输入，目标是预测其类。在学习了一组任务之后，在上对一个模型进行评估。

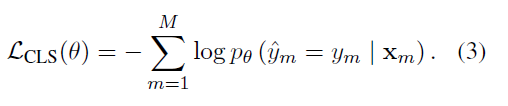
虽然我们提出的语言方法几乎适用于任何元学习框架，但我们使用的是原型网络（Snell等人，2017年），它对few-shot学习具有简单但强大的归纳偏见。原型网络学习一个嵌入函数例如，对n类的支持示例的嵌入进行平均以形成一个类原型（为了清楚起见，省略）：



给定一个查询示例，我们用与cn和之间的一些相似函数s概率比例来预测n类：



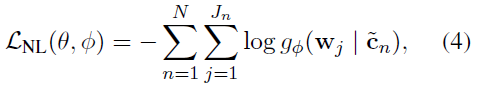
然后训练使分类损失最小化



**3.1 Shaping with language**

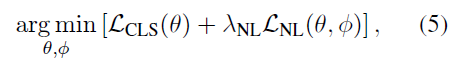
现在假设在训练期间，我们为每个类Sn都有一组Jn相关的自然语言描述。每个wj应该解释Sn的相关特征，而不需要与单个实例相关联。在图1中，我们有一个描述w1=（A；red；：：；square）。

我们的方法很简单：我们鼓励学习同样可以解码类语言描述的原型。假设是通过对n类的支持和查询示例求平均而形成的原型，然后定义一个语言模型（例如，递归神经网络），它定义在描述上提供了一个概率分布，并有相应的自然语言损失：



例如任务中所有类的类描述的总负对数可能性。由于依赖于参数，这个目标应该鼓励我们的模型更好地表示语言中表达的特征。

现在，我们共同将这两项损失降至最低：



其中，超参数控制自然语言损失的权重。在测试时，我们只需丢弃，并使用进行分类。我们称之为language-shaped learning（LSL；图1）。

**3.2 Relation to L3**

L3（Andreas et al.，2018）与LSL具有相同的基本组成部分，但定义的概念cn是语言描述本身的嵌入，由附加的递归神经网络（RNN）编码器生成。在训练过程中，地面真值描述用于分类，而则用于生成描述；在测试时，L3从中采样候选描述，根据相似度函数s使描述与支持集中的图像最相似（图1）。

与L3相比，LSL更简单，因为它（1）不需要额外的嵌入模块，（2）不需要测试时的语言采样过程。这也使得LSL在实际运行中比L3快得多：在没有语言机器的情况下，在我们的实验中，LSL在推理过程中比L3快50倍。

**4 Experiments**

这里我们描述我们的两个任务和模型。对于每个任务，我们评估LSL、L3和一个没有语言训练的原型网络基线（Meta；图1）。详情见附录A。

ShapeWorld。首先，我们使用Andreas等人使用的ShapeWorld（Kuhnle和Copestake，2017）数据集。（2018），包括9000个培训、1000个验证和4000个测试任务（图2）。每个任务包含一个K=4图像的支持集，表示一个视觉概念和一个相关（人工）英语语言描述，用概念的最小递归语义表示生成（Copestake等人，2016）。每个概念都是两个对象之间的空间关系，每个对象可以选择由颜色和/或形状限定，并且存在2-3个干扰形状。任务是预测查询图像x是否属于该概念。

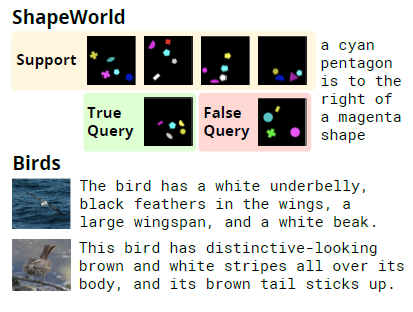
为了便于比较，我们使用与Andreas等人相同的模型报告结果。（2018），其中是通过两个完全连接的层馈送的固定图像网络预处理VGG-16（Simonyan和Zisserman，2015）的最终卷积层：



然而，由于固定图像网表示可能不是人工数据的最合适选择，我们还使用从头训练的卷积网络进行实验：或者是在few-shot文献（Chen等人，2019）中使用的4层卷积骨干网，如我们接下来描述的鸟类实验中使用的那样，或者更深层的ResNet-18（他等人，2016）。

这是few-shot学习框架的一个特殊的二元情形，只有一个正支持类S和原型c，因此，我们将相似函数定义为和正预测。是一个512维门控递归单元（GRU）RNN（Cho等人，2014），接受教师强迫训练。通过对验证集的网格搜索，我们设置了。

鸟。为了查看LSL是否能够扩展到更真实的场景，我们使用加州理工大学UCSD鸟类数据集（Wah等人，2011年），其中包含200种鸟类，每种都有40-60张图片，分成100列，50个验证和50个测试类。在训练过程中，通过从100个训练类中选择N个类对任务进行动态采样。然后从每个类中抽取K个支持和16个查询示例（类似于val和test）。对于语言，我们使用里德等人收集的描述。（2016年），AMT众工被要求详细描述鸟类的个体图像，而不涉及物种（图2）。



虽然每个图片有10个英文描述，但我们假设一个更现实的场景，其中我们只有在类级别上才有更少的可用语言：删除图片与其描述之间的关联，我们为每个类聚合D个描述，对于每个K-shot训练任务，我们从每个n类中抽取K个描述作为描述Wn。这使得学习对LSL来说特别具有挑战性，因为描述仅适用于单个图像的特征的标题会产生噪声。尽管如此，我们还是发现了一些改进，每个类只有很少的个描述，这是我们的主要结果，但也有不同的D，以查看模型使用语言的效率。

我们在N=5，K=1设置下进行评估，并使用Chen等人提出的4层卷积骨干作为。（2019年）。这里我们使用一个学习的双线性相似函数，其中W与模型联合学习。是一个200维的GRU，通过另一个网格搜索，我们设置了。

**5 Results**

结果见表1。对于ShapeWorld，LSL的表现比元学习基线（meta）高6.7%，至少和L3一样好；当从头开始训练时，表2显示了类似的趋势。对于鸟类来说，LSL比Meta增加了3.3%，但增幅较小，而L3低于基线。此外，LSL更有效地使用语言：图3显示了每类D的字幕从1（总共100个）增加到60（总共6000个）时鸟类的表现。LSL得益于非常少的标题，在过去20年中获得的收益有限；相比之下，L3需要更多的语言来平均接近基线性能。

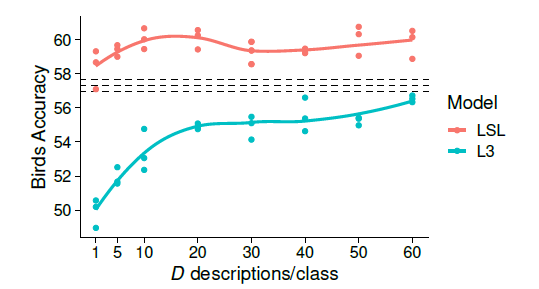


图3：改变每一类的描述，D，鸟类。每个点都是独立训练的模型。虚线表示独立训练的基线（Meta）。

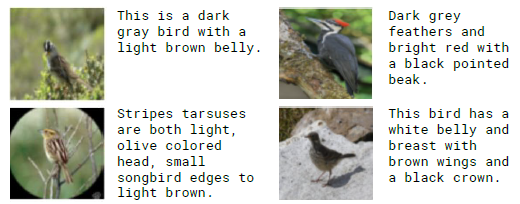
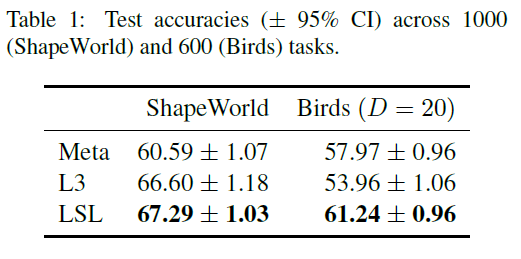


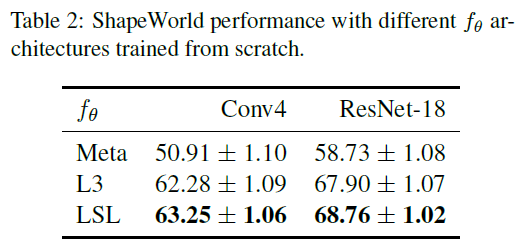
图4:L3解码器为鸟类验证图像生成的语言示例。由于LSL解码器是完全参数化的，所以它生成类似的语言。

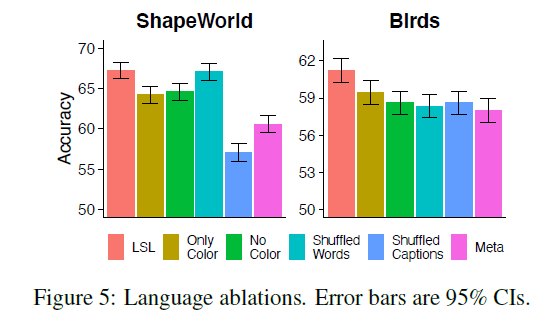
在低数据状态下，L3的低性能并不奇怪，因为它必须在测试时生成语言，而在数据如此少的情况下，这是很困难的。

图4中L3解码器的示例输出强调了这一事实：在某些情况下，语言看起来是合理的，但在另一些情况下，则存在事实错误（深灰色鸟；黑色尖嘴鸟）和流畅性问题。

这些结果表明，L3的任何好处都可能是由于语言对其嵌入模型的正则化作用，该模型已被训练用于预测语言的测试时间推断；事实上，离散瓶颈实际上在某些情况下是有害的。通过只使用正则化的视觉表示，而不完全依赖生成的语言，LSL是一个更简单、更有效和整体优越的模型。

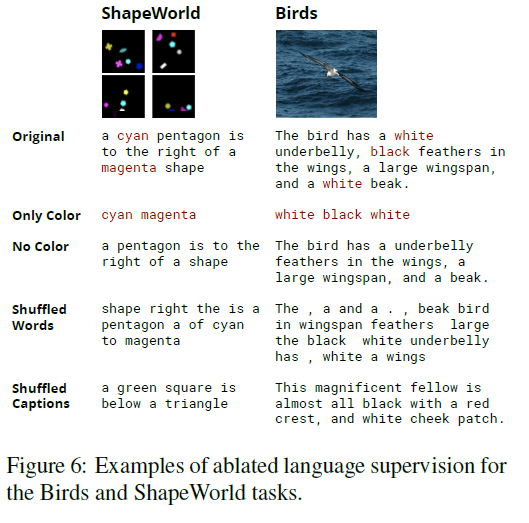






**5.1 Language ablation**

为了确定语言的哪些方面是最有用的，在图5中，我们检查了在烧蚀语言监督下的LSL性能：（1）只保留一个常用颜色词列表，（2）过滤掉颜色词，（3）在每个标题中洗牌，以及（4）在任务之间洗牌标题（参见图6的示例）。



我们发现，虽然颜色/无颜色语言的好处在不同的任务中有所不同，但两个组件都不提供完整语言的好处，这表明LSL同时利用了颜色和语言中描述的其他属性（例如大小、形状）。

语序对鸟类很重要，但对ShapeWorld却出人意料地不重要，这表明即使颜色和形状是分离的，模型通常也可以从示例中一致出现的形状推断出正确的关系。最后，当标题在任务间被洗牌时，鸟类的LSL并不比Meta差，而ShapeWorld则会受到影响，这表明语言对于ShapeWorld比细粒度的attributebase Birds任务更重要。

**6 Discussion**

提出了一种基于语言描述的镜头视觉识别模型LSL。LSL在两个任务中都优于基线，并且比L3更有效地使用语言监控。我们发现，如果一个模型被训练来揭示语言中的特征和抽象，那么在这些特征和抽象之上的一个语言瓶颈语言形状的表示是不必要的，至少对于这里探索的视觉任务来说是这样。

语言与足够丰富的属性和理论之间的界限是模糊的，最近的研究（Tokmakov等人，2019年）表明，通过对属性进行正则化，可以观察到类似的性能提升。然而，与属性不同的是，语言是（1）注释者更自然的媒介，（2）不需要对与任务相关的特征类型预先设定限制，（3）在无监督的形式中非常丰富。这使得用语言进行形状表示成为一种很有前途且容易获得的方法，以改进低数据环境下视觉模型的泛化。

**A Model and training details**

**A.1 ShapeWorld**

。就像安德烈亚斯等人。（2018年），首先从固定图像网预训练VGG-19网络的最后一个卷积层提取特征（Simonyan和Zisserman，2015年）。然后，这些4608-d嵌入被馈入两个完全连接的层，其中有一个ReLU非线性。

LSL。对于LSL，来自的512-d嵌入直接初始化GRU 的512-d隐藏状态。我们使用随机初始化的300-d字嵌入。使用GloVe初始化（Pennington等人，2014年）没有显著差异。

L3。和与LSL和Meta相同。是一个单向的1层GRU，其隐藏大小512与共享相同的字嵌入。最后一个隐藏状态的输出作为描述w（t）的嵌入。就像安德烈亚斯等人。（2018）测试时，每个任务共抽取10个描述。

训练。我们训练了50个阶段，每个阶段包括100个批次，每个批次有100个任务，使用Adam优化器（Kingma和Ba，2015），学习率为0:001。在训练过程中，我们选择了具有最高历元验证精度的模型。这与Andreas等人稍有不同。（2018年），每个车型使用不同的里程数，没有说明如何选择；否则，培训和评估过程相同。

数据。我们使用与Andreas等人相同的代码重新创建ShapeWorld数据集。（2018年），除了为更稳定的置信区间生成4倍的测试任务（4000对1000）。

请注意，L3和基线模型（Meta）的结果都比Andreas等人报告的分数低3-4分。（2018）（因为所有车型的性能都较低，我们对L3并不不公平）。这可能是由于PyTorch重新实现和/或使用更多测试任务重新创建数据集导致的模型初始化差异所致。

A.2 Birds

。4层卷积骨干与许多少数镜头文献中使用的骨干相同（Chen et al.，2019；Snell et al.，2017）。该模型有4个卷积块，每个卷积块由64个滤波器3x3卷积、批处理规范化、ReLU非线性和2x2最大池层组成。当输入图像大小为84 84时，这将导致1600-d图像嵌入。最后，用相似函数表示的双线性矩阵的维数为1600X1600。

LSL。生成的1600-d图像嵌入被馈入单个线性层，该线性层初始化GRU的200-d隐藏状态。我们用手套初始化嵌入。我们没有观察到通过增大解码器的大小而获得显著的增益。

三层。和是一样的。是一个单向GRU，隐藏大小200与共享相同的嵌入。最后一个隐藏状态作为概念。每个类在测试时抽取10个描述。我们没有观察到通过增加解码器g或编码器h的大小，或者在测试时增加每个类采样的描述的数量而获得的显著收益。

训练。为了便于与少数射击文献进行比较，我们使用与Chen等人相同的训练和评估过程。（2019年）。模型训练60000集，每集包含一个随机抽样任务，每个类16个查询图像。就像陈等人。（2019年），对600集进行评估。我们使用学习率为0.001的Adam，并在训练后选择验证精度最高的模型。

数据。就像陈等人。（2019），我们使用标准数据预处理和训练增强：imagnet平均像素归一化、随机裁剪、水平翻转和颜色抖动。