# 超过卷积的迭代视觉推理[[1]](#footnote-1)

赵淼译

(北京交通大学电子与信息工程学院，电子科学与技术，15650761201)

## 摘要

这里提出了一个新颖的迭代式视觉推理框架。这里解决了现在视觉系统缺乏推理能力的局面。这里有两个模块：一个空间记忆模块存储过去的通过并行更新的信息。一个全局的图推理模块。我们的图模块有三个部分：1.知识图谱（将类视作结点和建立边对不同类型的语义信息进行编码以获取他们之间的关系。2.一个区域的图模型将图像中的区域视为结点，将这些区域之间的关系视作边3.将图像区域与类别相对应。局部的模块和全局模块迭代运行和将彼此的信息交叉提供以优化结果。最后使用注意力机制将各个模块最好的结果进行组合形成最后的预测。该框架对缺失的区域具有推理能力。

## 1 介绍

这里认为现在的卷积网络缺乏推理能力，关键问题是对空间信和语义推理进行结合以构建下一代的视觉系统。

这里对于推理进行了两个轴的划分，一是通过推理获得空间信息的推理，语义上直接嫁接，另一个是语义上的推理，通过现有的信息关系推理出一个新的类别信息，然后在两个轴上进行融合，推理出其他位置的扩展的类别目标。



这里的问题是如何让推理更合理而不是离谱。

这里的关键是迭代的建立估计。最近有努力结合推理通过自上而下的模块，或是使用显式的记忆。前一种情况是使用基于类别的信息和低层次的特征联合提高识别的表现。后面的例子是通过序列化的目标检测，使用空间记忆存储检测到的目标，引导卷积网络提取密集的上下文信息以有利于后续的检测。

前面的方法有两个问题：1.使用堆叠的卷积块进行局部的像素级推理，缺乏全局的推理（使距离较远的信息也可以进行交流）。2. 两种方法都假设培训数据中有足够的关系示例 - 以便模型可以从头开始学习，但随着关系随着类数的增加而呈指数级增长，并不总是有足够的数据。很多语义推理需要从很少或没有例子中学习。因此，我们需要利用其他结构化信息进行视觉推理。

这里提出了一个通用的框架进行空间和语义推理。与仅依赖卷积的当前方法不同，我们的框架还可以以视觉识别知识库的形式从结构化信息中学习。我们的算法的核心包含两个模块：基于空间记忆的本地模块，使用ConvNets执行像素级推理。我们通过并行内存更新来提高效率。此外，我们引入了全局模块进行局域外的推理。在全局模块中，推理是基于图模型展开的。它有三个组成部分：a）一个知识图谱，我们把类当做结点，建立边来对它们之间不同类型的语义关系进行编码；b）一个当前图像的区域图，图中的区域是结点，区域间的空间关系是边；c）一个工作分配图，将区域分配给类别。利用这种结构的优势，我们开发了一个推理模型，专门用于在图中传递信息。局部模块和全局模块迭代工作，交叉互递预测结果来调整预期。需要注意的是，局部模块和全局模块不是分离的，对图像的深刻理解通常是先验的背景知识和对图像的具体观察间的折中。因此，我们用注意力机制联合两个模块，使模型在做最终预测时使用相关性最大的特征。

## 2 相关工作

视觉知识库：这里描述了前面的工作着重于知识的积累而不是知识的推理。

上下文建模：对上下文建模，对场景，目标和部分之间相互作用的获得是计算机视觉的核心问题。这里有场景推理，属性，结构化预测，关系图等不同角度来解决这一问题。这里卷积模块虽然没有显式的进行处理，但从局部到整体的聚合逐渐获得隐含的知识。最近自顶向下的模块，成对的模块，迭代式交互，注意力和记忆模块逐渐产生威力依赖于变异的卷积体。我们实现的是从结构化的视觉知识的基础上直接进行空间和语义关系的结合。

关系推理：早期是符号学的推理。基于统计学的模式提取。最近有许多人致力于使用神经网络进行结构化的数据图构建。尝试将输出规范化为一种关系和知识库的形式。在视觉信息下的这种尝试需要结合局部和全局的信息。

## 3 推理框架

这部分我们构建了推理框架。除了来自于卷积网络的简单预测p0，这里包含了两个核心模块进行预测。第一个是局部模块，使用空间记忆存储以前的信息并进行并行的更新。这时基于卷积范围内的推理。在卷积之上我们的关键贡献是实现了一个全局模块直接在区域和类表示为图中的结点。模块之间相互影响，交替进行估计。最后利用本地和全局的优势，我们将来自所有迭代的预测与一个注意力机制（Sec.3.4）结合起来，并对样本重新赋予权值（sec3.5）来训练模型（见图2）。



### 基于卷积的推理

我们第一个构建的模块称为局部模块，受4的启发。在高的层次上，这个主意是使用空间内存存储在已经找到的目标在他们已经找到的位置上。S是一个三维的向量。首先的两维H和W对应缩减到1/16的图像。第三维对应深度为512，存储每个位置潜在的有用信息。

S在高层次特征和中层次特征下更新。对于高层次，关于类标签的估计被存储。然而知道类别并不理想，更多的信息包括形状，姿势等等也是有用的。在这里在softmax之前使用了逻辑函数f，结合了在记忆中存储的底部卷积层的特征图。

对于给定的一个要更新的图像区域，首先对来自底层的特征图进行裁剪，缩放到预定义的大小（7\*7）通过双线性内插。因为高层特征是一个可以覆盖到所有位置的矢量，这里将其扩展为全部49个位置。两个1\*1的卷积被用来对信息进行混合和形成对于区域r的输入特征。记忆S中的相同区域也被裁剪缩放到7\*7，使用st指示。在联合后，我们用一个卷积的门循环单元写入记忆。

|  |  |
| --- | --- |
|  | 1 |

这里获得的是对于r更新的记忆，u是更新们，z是复位门，W和b是卷积的权重和偏移量，圆圈代表entry-wise乘积。更新之后的记忆通过剪切和缩放被放回了S中。

并行更新：上面的工作对记忆进行序列化的更新。然而这种形式对GPU不够高雄效。这里提出并行单元进行更新。在将区域返回S中时，我们也计算权重矩阵t，这里个矩阵的每个值是0到1之间表明一个一个区域对于记忆单元c有多少贡献，1表示对于区域的全覆盖，0表示对这个区域完全不涉及。最终更新值是所有单元的平均值。

实际中的推理模块包含卷积网络c带有3\*3的卷积和两个4096维全连接层，使用S作为输入，构建连接在局部感受野的窗口上进行预测。因为二维的架构的图像架构，位置信息在S中被保留了，这样的结构对于空间推理是十分有用的。

### 超越卷积

第二个模块在局部区域和卷积之上运行进行全局的推理。这里全局的含义是两层的，首先是空间上的，我们让各个区域直接交流信息，而不是限于推理模块中各自的感受野。第二是语义上的，者里使用了视觉知识库，他可以提供不同类图像上的全局的关系。为了获得两种类型的推理，我们构建了一个图G=（N，S）,N和S代表着结点的集合和边的集合，相对的。区域结点Nr对应于R区域，类结点Nc对应于类别。

对于s，在节点之间有三组边被定义。首先对于Nr，一个空间图对区域的空间关系进行编码。多重的边被设计去描述相对的位置。我们以基本的左右，上下的关系为开始和定义了测量像素级距离的权重。这里没有使用绝对的原始距离，而是使用正则化后的0，1之间的数值。使用了一个核函数，凭直觉认为更近的距离是更相关的。在邻接图中直接表示边的权值。另外对于覆盖模式进行了编码（IoU），这在描述两个区域之间的重叠关系时特别有用。这里会对一些错误的假设更有鲁棒性。

来自于知识库之间的语义关系被用来构建第三类边，在类别和类别之间的关系。这里包含多种类型的边。包括归类信息，整体与部分之间的关系，相似性的信息，这种是一些普遍真理和人们认为是常识性的知识。这些可以是手工罗列或是自动采集。有趣的是，这些关系可以极大的帮助识别。例如人骑车是一个关系，当对人和车的置信度较低时，这种关系就不太可能发生，这有助于去除一些虚假的联系。这里使用了两类知识图，一类构建在一些常识之上的边，另一类包含大规模累积下来的关系，实际参考4.1.

第二组边是区域和类别之间的边，这里对一个区域所属的类别进行了假定。同样的这里有多种类型的边。这样的边负担着传递信息的责任，包括区域到类别和类别到区域之间的信息。这里使用softmax的权值进行联系的指定，而不是单独的分类。

这样获得了基于图的推理模块R。输入到图里的是区域结点组合后的特征。Mr（R（R\*D维）。D是512是特征通道的数目。对于每个结点nc，我们选择了了off-the-shelf词向量来表示。（C\*D）维。这里对前面的工作进行了扩展。注意这里的目标是更好的识别区域，因此所有类结点应该被用更好的空间表示的踏板。伴随着观察，我们设计了两种推理路径学习输出的特征Gr：一是空间路径，这里只有区域结点。

|  |  |
| --- | --- |
|  | 1 |

这里Ae是边类型e的邻接矩阵，We是权重。第二个推理路径是语义上的推理路径在类节点之间：

|  |  |
| --- | --- |
|  | 1 |

这里首先通过，将区域映射到了类别上通过中间的类特征Mc，和再次聚合了来自类之间多种类型边的特征，最后区域Gr输出的是两个路径的结合。

|  |  |
| --- | --- |
|  | 1 |

这里使用非线性激活将语义信息传回到区域。

和卷积核类似，上述描述的路径被堆叠起来，输出Gr能够可以通过另一组图操作，允许框架表示联合的空间语义推理在深层特征下。我们使用了三个堆叠的操作在残差连接在R种，最后输出用来预测。

### 迭代的推理

推理的关键要素是要迭代的建立估计。信息如何在迭代之间交流呢。这里的答案是外显记忆，这里存储了过去迭代的全部历史。局部模块使用空间记忆，全局模块使用另一种记忆M不包括空间结构。在迭代i次下，Si使用卷积推理模块C为每个区域产生新的推理。相似的，全局模块从R种给出新的预测。这些新的预测作为高层次的特征可以用来更新记忆Si+1和Mi+1.这种新的记忆导致更新的fi+1，迭代继续。

虽然可以孤立的进行局部和全局的推理，但是这两个模块最好协同工作。因此，我们在生成预测时加入了完整的通道。在结束时，我们引入了交叉式信息连接。在推理之后，局部特征和全局特征被连接在一起更新全局特征用GRU。在这种方式下，空间记忆可以受益于全局性的空间和语义关系的知识，并且图可以得到一个更好的局部区域布局的感觉。

### 注意力

在注意力工作的启发下，这里对输出做出另一种修改，模型会产生表示现在的预测相对于之前迭代或模块的置信度的注意力值a。然后混合的输出是一个对于所有预测使用注意力进行加权的版本。如果模型迭代I次，输出包含2I+1（包括I个局部，I个全局和1个卷积网络）个预测，使用注意力an后，最后的输出可以通过下式计算：

|  |  |
| --- | --- |
|  | 1 |

注意这里fn式softmax之前的logit，然后被激活产生pn。这种注意力机制的引入允许模型智能的选择可行的来自不同模块和迭代上的表示。

### 训练

最后使用了一个端到端的网络结构，包括1卷积网络损失，2区域模块损失3全局模块损失4最后的预测损失

因为我们希望推理模块集中与更困难的样本，我对损失进行了简单的重新赋权，基于以前迭代的预测。形式上，对于在迭代i大于等于1时，全部模块的交叉熵损失按下式计算：

|  |  |
| --- | --- |
|  | 1 |
|  |  |
| 这里p票是对于真实分类的softmax输出，贝塔是0到1之间的树，控制权重分布的熵，贝塔为1是均匀分布，贝塔为0，熵最小化。在实验中，设置贝塔为0.5.pi-1作为特征不进行反向传播。对于局部和全局，p0r是卷积网络输出。  4 实验  数据集和图  这里对数据要求是紧密标记的，为了结合知识图谱，长尾分布的数据集是理想的测试标准。这里使用了ADE和基因计划数据集。  ADE使用了更为详细的分类。过滤掉了少于5个实例的类别，剩余1484个类。在数据集的部分标注的帮助下，一个常识知识图谱在5种类型的边的情况下被建立了。包括1.部分与整体2.类别指定3.单复数4水平对称性5.相似性。注意到前四种边是直接边，因此也包含其中的反向版本。  这里对于于VG是相似的预处理，使用了同义词集代替代替原始名称。知识图利用了数据集中的关系标注，自动选择了10种最常用的关系自动构建边。对每种类型的关系使用正则化的权重因此邻接矩阵每一列总和为1.这种方法构建的周四hi图更具噪声性，这允许我们证明我们的方法是鲁棒的。  在COCO上没有使用知识图，没进行测试集的分割，知识作为分析。  这里使用的任务是给定区域的预测，这里可以使用目标检测的分类器，但是这些目标检测会具有极大限制，丢失掉一些信息，因此可以验证模型的补充效果。  测试时使用平均准确率和分类精度做测试，不需要IoU。计算一个集合中每类的准确率，然后取平均。对一些罕见的类分配更大的权重。、  实施细节  简化版本的tf-faster-rcnn被用来作为实施基准。这里预训练的ResNet-50用作图像分类器，图像被扩大到短边600.整幅图像共享卷积特征图，conv4作为计算层。Grougtruth boxes 用作感兴趣的区域计算区域特定的特征（裁剪和缩放到7\*7）。Conv5之后的所有层被调整获得最后基准特征预测p0.批归一化参数被固定。  对于局部模块，使用conv4作为终极特征放到空间记忆S中。对于全局模块，中间特征是conv5的2048维特征在经过平均池化后。两个特征使用softmax f作为混合。词向量来自fastText被用来表示每一类，这里提取各个子词的信息作为到词典外单词。ReLu作为激活函数。推理模块将重复三次和每次更新所有区域，更多的迭代不会提供帮助。  5\*10-4作为学习率，动量0.9.使用320k次迭代。训练时使用左右旋转。  分析  对于难例赋予更高的权重可以提升大约0.5. |
| 这里p票是对于真实分类的softmax输出，贝塔是0到1之间的树，控制权重分布的熵，贝塔为1是均匀分布，贝塔为0，熵最小化。在实验中，设置贝塔为0.5.pi-1作为特征不进行反向传播。对于局部和全局，p0r是卷积网络输出。 4 实验数据集和图 这里对数据要求是紧密标记的，为了结合知识图谱，长尾分布的数据集是理想的测试标准。这里使用了ADE和基因计划数据集。  ADE使用了更为详细的分类。过滤掉了少于5个实例的类别，剩余1484个类。在数据集的部分标注的帮助下，一个常识知识图谱在5种类型的边的情况下被建立了。包括1.部分与整体2.类别指定3.单复数4水平对称性5.相似性。注意到前四种边是直接边，因此也包含其中的反向版本。  这里对于于VG是相似的预处理，使用了同义词集代替代替原始名称。知识图利用了数据集中的关系标注，自动选择了10种最常用的关系自动构建边。对每种类型的关系使用正则化的权重因此邻接矩阵每一列总和为1.这种方法构建的周四hi图更具噪声性，这允许我们证明我们的方法是鲁棒的。  在COCO上没有使用知识图，没进行测试集的分割，知识作为分析。  这里使用的任务是给定区域的预测，这里可以使用目标检测的分类器，但是这些目标检测会具有极大限制，丢失掉一些信息，因此可以验证模型的补充效果。  测试时使用平均准确率和分类精度做测试，不需要IoU。计算一个集合中每类的准确率，然后取平均。对一些罕见的类分配更大的权重。、 实施细节 简化版本的tf-faster-rcnn被用来作为实施基准。这里预训练的ResNet-50用作图像分类器，图像被扩大到短边600.整幅图像共享卷积特征图，conv4作为计算层。Grougtruth boxes 用作感兴趣的区域计算区域特定的特征（裁剪和缩放到7\*7）。Conv5之后的所有层被调整获得最后基准特征预测p0.批归一化参数被固定。  对于局部模块，使用conv4作为终极特征放到空间记忆S中。对于全局模块，中间特征是conv5的2048维特征在经过平均池化后。两个特征使用softmax f作为混合。词向量来自fastText被用来表示每一类，这里提取各个子词的信息作为到词典外单词。ReLu作为激活函数。推理模块将重复三次和每次更新所有区域，更多的迭代不会提供帮助。  5\*10-4作为学习率，动量0.9.使用320k次迭代。训练时使用左右旋转。 分析 对于难例赋予更高的权重可以提升大约0.5。 |

空间记忆S对于局部模块是极为重要的。直接取代这一模块使用conv4输出大幅度降低模型表现。局部背景向量C影响不大。在全局模块上相反，移除推理模块打幅度降低模型表现，远超过记忆模块M。最后移除了交叉信息提供和降低了迭代次数导致表现更差。



缺失区域：这里使用Faster RCNN极大值抑制的方法过滤掉了一些ground truth，这里并未使用Faster RCNN提出的区域方案，而是使用已知的区域ground truth。一是在推理前过滤掉一部分区域，而是在推理后过滤掉一部份区域。这两者在表现上没有明显区别，这里缺失了知识图，全局模块比局部模块工作的更好。

当区域被丢弃的时候，推理模块的性能下降平滑。并且可以认为1区域建议通常与简单的例子相对应，在这些例子中，需要更少的上下文，而2像我们这样的上下文推理框架从更多已知的区域中获益。当区域缺失过于严重时推理会导致性能的破坏。





1. 赵淼 电信研1706班 电子科学与技术 17120036 15650761201 [↑](#footnote-ref-1)