Spatial Memory for Context Reasoning in Object Detection阅读笔记

## 简介

对实例级上下文和目标-目标关系建模是具有挑战性的，现有的目标检测技术并不具备任何记忆功能，其并行地检测所有目标，并通过non-maximal suppression（NMS非极大值抑制）避免对目标的重复检测。考虑到建模目标之间的关系，需要利用到空间上的推理，不仅要记忆空间布局，还需要建立有效地推断模型来提取空间信息，作者提出了Spatial Memory Network（SMN）。其本质上将目标实例装回伪图像表示，这样可以简单地喂给CNN实现目标间上下文推理。

上下文的常见建模方法有两种，一种是图像场景级的建模，另一种则是对象间关系的实例级建模。尽管CNN已经取得了不错的表现，在隐含层中似乎也能够实现上下文的推断，但作者认为使用独立地空间记忆对于目标检测相关任务任务来说仍是接下来重要的一步，理由如下

1. 当前基于region的目标检测技术仍旧把其看作一个感知问题而不是推理问题，分类器会在目标附近生成多个检测结果，并且依赖手工设计的具有为de-duplication（重复数据删除）预定义阈值的NMS。这个过程是次优的，而对于空间记忆来说，其将已检测的目标记忆化，可以自动学习到NMS的功能。
2. 考虑到空间记忆支持存储语义和位置信息，下一步就应该做完整上下文的推理，比如根据现有布局推断其他实例是什么，在哪里。
3. 空间记忆本质上是对实例级视觉知识的编码，需要正确处理目标群体间的空间和语义关联。

## Faster RCNN

作者的空间记忆网络不关心目标检测模型是什么，本文中作者使用FRCNN。具体结构参考Faster-RCNN的学习笔记。

#### 基础网络

使用VGG16作为特征提取，图像长宽各缩减为原先的1/16，再使用FRCNN进行region proposal和region classification。

#### region proposal

其本质上训练的是类别无关的对象性分类器，从滑动窗口中找到前景对象。

#### Region Classification

因为生成的提议候选框ROI可能是任意大小的矩形，因此使用ROI pool替代传统pool，使得输出尺寸保持固定。1000-way的全连接层被替换为两个全连接层，分别为C-way分类和边界框回归器，C个类别中的每一个都具有独立地边界框回归器。

#### De-duplication

注意到NMS在区域推荐和最终的检测环节都产生作用，作者考虑在区域推荐的时候依旧使用NMS来处理冗余的候选框，而在最终检测步骤使用SMN帮助网络选择合适的区域（具体见下文设计的新Loss）。

在此处，作者依旧运用NMS来修建可能对应于同一个物体的ROIs，但对于每一个类，作者关注于建立模型来替代per-class NMS，编码类间相互关系以抑制冗余的检测。这也是本文的重点。

## Spatial Memory Network

对于目标检测，目的是对于给定图检测所有的实例。代表最大化对数似然估计

这样的概率分解是严格的，无所谓选取的顺序。对于region-based目标检测器，有如下近似

NMS用于对检测序列的相关性进行建模，因为NMS决定于重叠的部分，相对于更具局限性。因此作者希望对编码到空间记忆中，具体一点，引入记忆变量，每当一个目标被检测到，则更新。

其中随着一同优化。

#### Memory

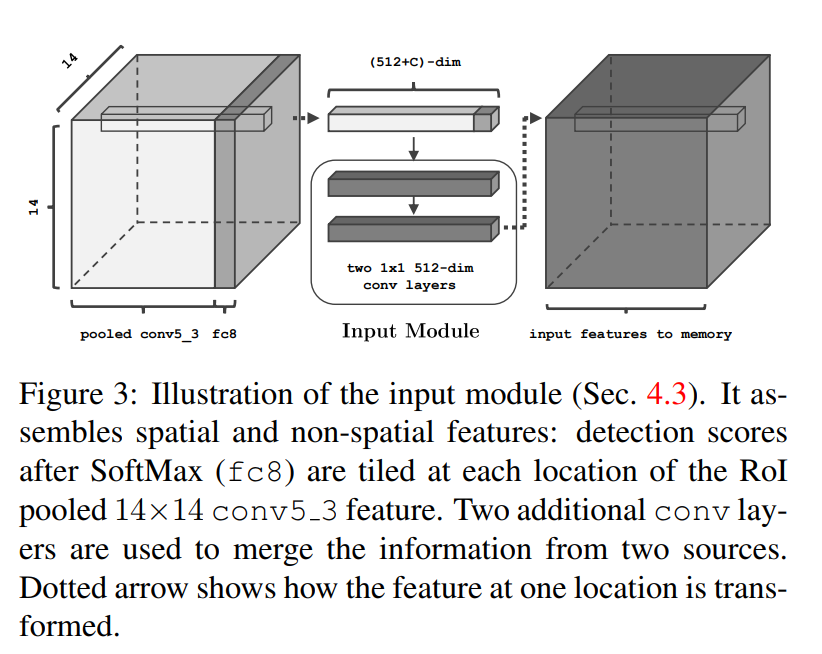
作者为图像任务设计了二维记忆，一方面因为图像是二维的，另一方面方便引入CNN来实现上下文推断。因为FRCNN的conv5\_3已经可以捕获很小的目标，作者设计二位记忆大小为，每个位置都是256维向量，其包含了至今提取到的特征。理想情况，最初记忆应该能够捕获自然图像在摄影上的偏差（比如太阳总是在上半部），但这种先验不应该决定于图像尺寸，所以每次会将其resize（双线性插值）以匹配conv5\_3的输入特征大小。

#### Index

现有方法往往面临记忆cell和输入内容的关联问题，如何实现索引。在作者的设计下，2D图像和2D记忆可以比较简单地关联。对于尺寸不同的输入框，利用ROI池化也可以精准的操纵需要从空间记忆中读取的内容。剩下我们就需要实现对记忆的写操作以实现记忆更新。

#### Input Features

这部分考虑，什么特征应该引入记忆。conv5 3提取到了图像的空间信息（通过不取Max的ROI池化后resize到）。不过这还不够，我们还需要关于检测的分类信息，当两个目标在同一个区域出现时，这个操作就尤为重要。因此作者把fc8的softmax得分也作为输入加入在conv5 3输出的每一个位置之后，并用两个卷积进行整合。这里采用得分而不是最终的分类one-hot向量，因为这对失败检测（即什么目标都没有）的鲁棒性更强。



#### Writing

在给定区域和输入特征时，用convolutional Gated Recurrent Unit更新记忆，使用卷积替代全连接层，GRU具有reset gate和update gate，每个区域位置共享，用Sigmoid激活。Tanh用来约束记忆值保持在。

这个过程即从取出的region先被ROI池化成，经过GRU后通过逆ROI操作复原并放回。

#### Context Model

得到编码的记忆后，用在顶端堆叠卷积网络的方式实现上下文推理。使用5层全卷积网络提取空间pattern，每一个卷积核为，并用padding保持m-conv5和conv5\_3大小一致。另外每两层建立一条残差连接。

#### Output

在FRCNN中用同样的方式对待m-conv5，三个卷积层选取推荐区域，两个全连接层（各2048个单元）和ROI池化实现区域分类。换句话说，我们把conv5\_3（来自预训练模型的特征）和m-conv5（来自记忆网络的特征）都用FRCNN中相同的结构处理（参数独立），并在最后使用全连接层对两者的输出进行整合。

作者设计了另一种残差结构来结合记忆得分和FRCNN得分：当记忆为空时直接使用FRCNN来检测，从第二次迭代开始，将记忆预测放在FRCNN上部，这样记忆可以提供额外的上下文。此结构除了便于可视化上下文引入带来的变化，对整个网络收敛也非常关键。

#### Selecting Next Region

空间记忆的引入让目标检测变成一个序列问题，那么如何选择下一个区域？作者提出一个贪心的方法，选择置信度最高的前景目标box。

## Training the Spatial Memory

网络用BPTT训练，更新权重时，将网络进行多次展开。不过除了梯度消失问题，在目标检测任务上强制增加条件结构也对训练造成一定影响。作者实践中最严峻的问题是deduplication。

#### Learning De-duplication

作者尝试了四种不同的架构，其中前两种均无法收敛，究其原因是网络持续接受两种矛盾的信号：其中一种引导感知，另一种对抗信号在阻碍更多的感知。我们直觉上希望网络避免对已检测到的物体重复检测，但FRCNN则引导网络做更多的检测。由于也是从头开始训练，其提供的信号弱且不可靠，这两种信号反馈给导致学习困难。

为此作者选择从第二次迭代开始截断对FRCNN的梯度，网络可以正常收敛。作者进一步把第一次迭代中和空间记忆有关的权重删去，最终得到的结构便于观测引入记忆带来的变化并且可以单独训练。

#### RoI Sampling

因为空间记忆学会了去重复，那么引入一个现象，就是刚刚检测到的物体，下一次迭代，其区域的前景背景标签会反转。我们希望这些区域能起作用，抑制重复检测，所以为了防止这些区域被埋没在大量负区域中，引入RoI采样。RoI采样的特性是只有少量的区域最后被采纳，往往置信度最高的区域不一定被选中用于更新记忆。这给了我们嵌入其他高可信box的机会（包含标签翻转区域的机会），可以抑制过拟合。

#### Multi-Tasking

多任务的动机在于，现有的任务只是寻找目标，并不一定能转化为对先前检测目标的记忆，最终可能只记住了它的位置而不知道其类别，换言之空间记忆网络的收敛效果并不好。为此，作者引入一个重建Loss，让网络额外预测目标在记忆中的类别。作者在m-conv5顶部增加了一组和原网络一样的分支作为FRCNN，每一次迭代中同时执行区域推荐和区域分类。这些权重只用于预测之前检测到的物体。通过强迫从记忆网络检测出以前的物体和其类别，进一步优化空间记忆。

#### Stage-wise Training

得益于作者的架构，模型可以通过两个独立的分段进行训练。FRCNN ,这一检测模型可以先单独训练。带有空间记忆的推理模型在优化后的顶部进一步学习。这样可以孤立基本模型的影响，专注于训练空间记忆网络。

作者采用了课程学习策略（即先用简单任务训练，再逐步提高难度，直到和目标任务匹配，以达到较好的收敛效果）：让迭代次数较少的SMN引导更多迭代次数的SMN。这可以让网络更关注于目标之间的关系。

## 总结

作者引入了空间记忆网络来高效有效地对实例级别地上下文进行建模，其本质上将图像实例装入了伪图像表示中，这使得其易于和卷积网络对接以提取上下文信息执行对象关系推理。这个结构的引入使得目标检测相关任务的准确性得以提升。