Iterative Visual Reasoning Beyond Convolutions阅读笔记

## 简介

现有的推理模型一类是使用至上而下的模型，组合高级特征和低级特征，提高识别性能；另一种是引入额外的记忆。其均存在两个问题：堆栈式的网络堆叠对局部像素推理较为有效，但全局推理能力弱，各区域无法实现有效地信息交流。且这两类模型都需要足够的数据提取关系，而实际上可能只存在少量或者不存在数据的关系等待提取。

为了实现更加深层次的图像知识推理（空间上、语义上以及两者共同作用），作者提出了一个通用框架。其包含两部分：

* 基于空间记忆的局部模型，利用CNN实现像素级推理。
* 引入全局推理模型，其基于图结构，具有三个组件。
  + 知识图谱，将类别表示为节点并建立边以编码不同语义关系的类别。
  + 区域图谱，当前图片的区域图谱中各区域是节点，区域之间的空间关系是边。
  + 派发图谱，将区域分配给各个分类。

## 推理框架

推理模型具有两个核心依据，前者为基于记忆的局部模型，后者为全局的区域-分类关系模型。通过引入attention机制，组合迭代中来自两者的预测。对于困难样本，使用sample re-weighting进一步训练。

#### Reasoning with Convolutions

此模组定义为使用卷积网络从空间记忆中提取局部关系。

空间记忆网络储存了之前检测到的目标被保留空间关系。特征抽取后，长宽分别缩减为原图的，特征深度。记忆中的每一个cell都是一条长度为的向量，存储像素级的隐含信息。

通过高级和中级特征更新，高级特征指分类标签信息，中级特征指物体的形状动作等等。在本文中，作者在softmax之前使用logits函数将概率映射到。结合来自卷积层的特征，喂给记忆网络。

对于区域，对应的下层特征区域被切出来并resize到固定的方块作为。因为高级特征的向量是对整个区域的概率表示，所以将其内容添加到每个特征向量尾部，再用两个卷积整合。最后再用GRU写入记忆。

其中是被区域更新的记忆，是gate，和以及是卷积参数，代表entry-wise乘法，表示激活函数。

#### Beyond Convolutions

此模组定义为，其超越了局部区域和卷积，实现全局推理。全局有两层意思，其一为空间，希望各区域能有效地传递信息而不局限于感受野；其二为语义，希望从知识库中受益，提供图片中各类别的关系。为此作者构建了一个图网络，其中表示点集和边集。

点集有两种类型组成：表示区域，表示类别。

边集有三组类型：

* ，用于编码空间关系网络，其中分为left/right,top/buttom关系，并且用权重表示距离，距离通过规范化到。此外边还被用于编码覆盖的pattern，表示区域覆盖。
* ，表示目标出现的位置，其需要具备双向推导。作者考虑使用完整的soft-max分数，将每个区域和所有类别相连，分数作为权重。这样提取的信息更多，对失败区域的鲁棒性更强。
* ，来自知识库，约束类间关系。其关系有多种，如is-kind-of,is-part-of,similarity。这些人类常识可以通过手工方法获得也可以自动获取（参考NEIL）。

将推理模型表示为，表示从所有区域节点组合而来的特征。对于作者采用现成的词向量模型作为类型表示。

作者构建了如下推理路径以学习输出特征

是边类型的邻接矩阵，是用于对输入进行变形的权重。

将区域映射为类，将中间特征整合类特征。

最终表示为

将语义信息映射回区域，再使用非线性激活。

#### Iterative Reasoning

迭代过程中如何实现信息的继承，作者采用类似空间记忆的方式，为全局模块也设计了记忆，其剔除了空间结构。回顾空间记忆，在迭代中在卷积推理模型后对每个区域生成新的预测，此网络也从中输出预测。预测作为高位特征可以用于更新。新的记忆则会在新一轮中更新。

尽管两个模组可以独立运作，若能联合运作再好不过，作者引入了cross-feed连接，在推理之后，局部特征和全局特征被拼接在一起，使用GRU更新。

#### Attention

作者引入了Attention机制，在生成预测之外，额外预测了attention值，表示当前预测与其他迭代或模块预测的相对置信度。最后整合的输出是利用将各预测带权求和。

此时的依旧是logits，经过softmax后得到预测概率，Attention机制使得模型可以从各迭代轮次、模组中智能选择好的预测。

#### Training

最终框架可以端到端地训练，loss function由三部分组成

1. 基本的卷积loss
2. 局部模组loss
3. 全局模组loss
4. 最终基于Attentions的预测损失

作者希望模型能更关注于困难的样本，所以作者提出对loss加权。

这一项可以让难以区分的区域获取更高的loss权值。

## 总结

作者在使用记忆模型的Faster-RCNN之上，引入了可迭代推理的图结构以编码空间和语义关系，并使得信息可以在图网络上传递。额外的组件使得目标检测任务不再局限于感受野内做像素级的检测，利用到了regions之间的诸多信息进行推理，提高了目标检测的准确率。