# Visual Knowledge Learning阅读笔记

## 总览

当今的发展状态下，AI系统尽管以及能够完成一些工作，但在理解图像这类更深层次的工作中，其需要非常大量的知识规则，而我们没法手工提供。

传统方法中，我们手动地总结这些知识，转化成结构化的数据，在任务运用中仍遇到两大难题：

1. 尺度：即使像ImageNet这样的大型图像数据库，其涵盖的图像数量(1 million)仍然比不上文本知识库的规模(1.6 billion)。此外对于图像的标注，完全跟不上网络图像的增长速度，这限制了传统方法对知识理解深度和广度。
2. 可用性：文字知识可运用于开放性的问答系统中，然而想要把一些事实运用到图像理解中则更为困难。根本上，稀疏的干净的知识条目和稠密的杂乱的视觉世界之间存在鸿沟。传统的结构化知识库很难独立解决视觉问题。

本文着眼于解决上述两个问题。首先，不同于手动标记万物，通过构建一个学习系统自动地学习视觉知识。其次，在具备这些知识后，让它们为高阶任务提供益处。

#### 视觉知识

本文所关注的知识可分为显式(explicit)和隐式(implicit)。

显式知识是可以表达为一个规范形式的，比如主语-动词-对象三元组，或者被标记的类型实体。

隐式知识往往难以表达，但对于技能的获取和端任务的完成至关重要，称作常识一词更为贴切。因为其天然性，人们很少交流这些内容，但其又是人与人之间连接的共同点。

#### 视觉知识的学习

对于显式知识的学习，作者将互联网作为真实世界的代理，让机器自动从网络图像中学习。

作者提出了NEIL系统自动地从网络中提取显式知识。

在此基础上，作者将先验知识运用于前景分隔任务中。

随后作者在NEIL基础上解决分类歧义问题。

#### 视觉常识的学习

不同于显式知识，常识并不是随处可见的。不过也许我们不需要显式地得到它们，最终我们总希望机器能获取完成任务的技能，而知识往往作为中间环节帮助优化目标，因此我们可以考虑让隐式知识蕴含在模型参数中。

作者从模型中间层的输出以及预训练模型角度去理解知识的融入，即使网络图片存在较多的噪音，但是其规模远大于现有的数据库，其能够提高模型的性能。

#### 视觉知识处理

隐式表示在端到端训练中有显著的作用，稠密的向量化的表示缺失可解释性，但易于运用和迁移。另一方面，显式知识用一种直接的方式解释自身，能够给模型运作给予一个明确的解释。

同时作者认为在无监督学习发展不足的现状下，隐式知识的运用需要显式知识的的帮助才能奏效。

作者提出了空间记忆的概念，考虑在卷积网络中以空间角度进行上下文推导。但其不具备处理结构化的知识的能力，而且推导只局限在图片的局部区域，这与人将现象和背景建立关联的能力相距甚远。为此，作者提出了一个基于图的模型，通过传递消息来进行推理。

#### Learning Explicit Visual Knowledge

本章主要关注于自动地提取结构化的视觉知识。

###### Never Ending Image Learner

本文作者设计了一个利用名词短语在搜索引擎收集图片，并从中提取结构化关系的模型。其特点在于收集关系的过程是半监督的，利用搜索引擎提供文字图像对应关系作为seeds，训练基础的检测器，利用它们进一步挖掘图像中的关系。

目标、属性、场景之间，利用在图像中的共现关系提取结构化的关系表示，其可以作为约束进一步优化未来关系提取的准确性。

###### From Bounding Boxes to Segmentations

本文同样利用了搜索引擎，设计了可以自动得到物体前景分割的模型。利用搜索引擎结果背景单纯、目标居中的特点，再使用图像的HOG特征得到初步的聚类簇（具有视觉同质的特点，同一聚类目标大多位于图像中央且外形轮廓基本一致），使用graph-cut进行前景分割，此时的分割结果还不能作为知识保留下来，因为其存在较多的噪声。利用现有簇的检测器对图片进行打分，再进一步合并聚类得到视觉子类簇（期间噪声数据点会被丢弃掉，并且合并后的视觉子类依旧可以保持视觉同质的特点）。利用视觉子类良好的对齐特点，将已有样本的前景分割掩码作为graph-cut算法的先验知识，为后续分割提供帮助，并不断扩充先验。

此过程也是半监督的，其得到的前景分割可以为视觉知识库扩充内容。

###### Sense Discovery

本文提出了视觉特征和文字特征新的映射关系，设计了一个图模型编码视觉特征和文字特征聚类簇之间的关系。

图模型中，存在若干表示文字聚类和图像聚类的节点，两者之间建立无向连接。作者设计了图的得分以优化图结构。约束了特征相似的数据点应该分配到相同的聚类簇中，此外在文字域上相似的数据点在视觉域上也应该具有相同的聚类。

通过迭代优化的方式，不断改进图的结构。

#### Learning Implicit Visual Knowledge

本章探究的是学习隐性的视觉知识，因为无法像显性知识那样结构化地表示，本章引入了神经网络实现对隐性视觉知识地学习和编码。

###### Using Query for Supervision

本文作者引入两步训练法，先利用较为纯净的图片训练基础的网络，易于网络学习到目标的关键特征。此时还可以通过网络的输出结果，提取出不同语义概念之间在出现概率上的先验关系。

再通过更复杂语义更丰富的图像资源在关系约束下进一步训练网络，使得网络不仅具备提取目标特征的能力，还隐含了丰富的语义信息。

###### Image to Caption and Back

本文探究的是图像与文字描述的双向关系

模型需要实现

* 已知文字序列输出视觉表示
* 已知视觉表示输出文字序列

作者使用了RNN模型，由于RNN对序列远端的感知能力较弱，作者提出了引入一个隐含层来表示文字序列对图像的理解，在上述两个任务中，隐含层作为先验既可以帮助重建视觉特征，也可以帮助预测文字序列的下一个词语选取。

#### Reasoning with Visual Knowledge

本章节主要关注视觉知识的推理，将捕获的视觉知识运用于提升端任务的性能。

###### Spatial Memory Network

传统目标检测中，目标与目标之间的检测是相互独立的，不存在信息的共享。作者考虑将已被检测的目标的空间和分类信息作为先验知识，帮助检测下一目标。

本文通过在Faster-RCNN的基础上引入空间记忆网络，在检测的过程中将被检测到的目标的视觉特征和分类特征写入记忆网络中空间对应的区域。在下一轮检测时，记忆网络就可以为目标检测提供额外的得分。

###### Iterative Reasoning

本文引入了两个模块，空间记忆模块捕获图像的局部知识，全局推理模型用以编码目标间的空间关系、目标区域和分类的关系、类间关系。在检测过程中，利用迭代推理的方式，不断更新两个模块的信息。此外作者引入Attentions机制，对区域的预测结果进行加权。此模型更加有效地利用了图像的深度信息，提高目标检测任务的准确率。

## 总结

本文主要解决了视觉知识处理中的两大问题：

* 视觉知识的自动抽取
* 利用视觉知识进行推断

除此之外，作者提出了一些有待探索的关系：

* 身体关系：引入关于人体的一些常识。
* 功能关系：考虑物体在功能方面的属性。
* 时间关系：现有的方法都只关注与静态的图片，而一些知识会与时间有关，可能需要通过视频或者相册来学习。

关于视觉知识。显性知识具有良好的可解释性，因此可以为隐性知识的提取提供帮助，尤其是在无监督学习效果不佳的现状下。隐性知识不仅需要使用于当前的任务，还需要具有较好的迁移性。如何将两者互相结合是未来探究的方向之一。

在视觉知识之上，可能还可以探索更深层的关系，比如数学、物理上的规律。CNN是良好的函数拟合工具，也许可以通过模型理解这些世界的规律。

现有的推理模型还是在非常有限的环境下进行测试的，未来需要尝试将它们运用到更加实际的环境中。

关于至上而下和至下而上。本文中，作者在分割任务使用了自上而下的信息（前进分割掩码）和自下而上（从特征抽取和视觉聚类到生成前最后的景分割）的方法。在空间记忆网络中作者也引入了自上而下的信息（从检测结果和预测结果向下传递）。这两种方法都进一步推进了对图像的理解。

关于端到端的学习。从手动提取特征到自动学习表示，这个过程中我们关注到的问题就是端到端的思想。一些基于统计信息所捕获的关系，不一定能适应端任务。未来在端到端的模型设计上也有更多的探索空间。