# Webly Supervised Learning of Convolutional Networks阅读笔记

## 简介

为了利用网络上的图像训练CNN网络，作者提出了两步训练法。首先使用简单的图像训练一个初始的视觉表示，利用这个初始的CNN通过引入结构化的数据和分类信息，将其应用到更困难的图像中。

## 背景

现阶段利用人类手动的对图像标记的耗费很大，且存在诸多问题。因此更好的办法是从不使用任何手工标记的网络数据中直接学习。现阶段有这么一些问题值得思考：

* 什么是好的数据源？现阶段研究者尝试从各种搜索引擎中获取数据。
* 什么类型的数据值得开发？例如图片、文字、图文混合、图片和n-gram模型。
* 如何开发数据？概率模型、exemplar based models（基于范例的模型）、deformable part model（可变性零件模型）、self organizing map（自组织映射）
* 我们需要从网络数据中学到什么？现有很多尝试在于训练视觉模型，或者从数据中探索常识关系。

然而，这些方法即使输入了非常大量的图像，其表现依旧劣于同期的接收人类监督图像的方法。原因如下

1. 网络上的无监督数据存在噪声。
2. 搜索引擎得到的数据存在偏好，搜索引擎运行在高精度、低召回的模式，倾向于目标出现在中心且背景干净、视角规范的图像。

除了数据外，人们所使用的算法中，传统算法如HOG依赖于手工特征提取，再如SVM只包含少量的参数，无法充分利用大量图像。基于记忆的最邻近分类器能很好捕获大量数据的分布，但对噪声的鲁棒性不足。CNN模型能够利用大量图像充分学习，尽管一些任务无法得到最好的效果，但易于迁移到其他相关任务中。

## 方法

作者的目标是从大量网络图像中学习到深度表示。CNN需要提供大量数据，小数据容易引起过拟合。此外作者发现利用随机的图像-文字/Tag pair依然很困难。比如Flickr的tag往往对应于元信息和特定地点，这会导致同标签内的大幅变动。一种解决方法是利用商业文字搜索引擎以提高训练集的多样性。但查询请求的文字，一部分可能并不对应某个具体的视觉概念，一些则过于具体（比如抽象的概念和随机的人名，所得到的内容的多样性区别很大）。这种不可视的场景和细粒化的分类会在训练时引入不可预见的噪声。

作者引入了两部训练方法来训练CNN。首先，模型学习一些简单的样本（居于中心、背景纯净）。然后一步步适配到困难样本中。考虑到Google靠前的结果往往利于学习（背景更干净，目标更居中）。一旦获取了表示，则使用来自Flickr的较为困难的图像，此时训练Flickr的图像依旧很困难，因为label中的噪声。为此，作者在利用Flickr做fine-tuning时引入约束。约束基于不同分类之间相似度关系，更明确一点，在学习简单图形的同时，建立一个关系图，在fine-tuning时，误差会通过图反向传播起到正则化的作用。

#### 网络初始化

常见的分类可以被良好地学习到，搜索引擎也可以给出相关的纯净结果。因此，不再使用随机的名词短语而是从ImageNet Challenge、SUN database、NEIL knowledge base中生成三个目录列表。ImageNet的同义词集通过采用第一个解释来转化成表面形式，他们大多数都关注于对象类别上。为了降低噪音，移除SUN categories名词前缀（往往是一些属性修饰词）。因为NEIL是面向搜索引擎设计的，所以它的类别列表易于理解。列表中的类别名直接用于搜索引擎查找图片，除了删除不可读的一场图片，不需要做额外的预处理。每次查询反馈了约600张图片，这些图片将直接喂给CNN模型。

为了公平起见，作者使用了和BLVC引用的一样的网络结构（除了输出层）。其由5层卷积层和两层全连接层堆叠，最后还有额外的一层全连接层用于预测类别标签。

#### 特征表示和图的适配

上文的网络在第一步收敛后，其已经学习到了一个低级滤波器，用以表示由Google搜索刻画的视觉世界。正如之前提到，现在的表示是偏差的，偏向于干净简单的图片（引擎反馈的车的图片有40%是45°拍摄视角）。而且，当某个概念是产品，它的许多图片都是壁纸或者人工背景的广告，以产品为中心，以最易销售的角度刻画。另一方面，照片分享网站如Flickr具有更多真实图片，尽管偏差依旧存在，它们和人类真实生活的视觉世界已经很接近了。因此下一步以Flickr的图像进行fine-tuning。

因为采用了Flickr的图像，其具有更多的噪声样本，强如CNN对噪声也会受到不小的影响。人类之所以能抵抗噪声，因为人不是在独立地识别某个概念，而是建立起了事物的彼此连接并发展出了理论以更好地理解世界。因此建立一个关系网络可以为识别训练提供更多信息。比如iPhone，Google提供的多是产品信息，Flickr提供的多为由iPhone拍摄的图像。

作者提出了一个数据驱动的方法，用训练集数据测试网络，取confusion matrix（混淆矩阵）作为关系，关系定义为：

其中是属于类别的图片集，返回集合大小。表示给定图像，其属于类别的概率，为了让网络更稀疏，只选择前关系最紧密的前个，并对probability mass（概率质量，离散空间中使用，对应连续空间的概率密度）进行归一化。

构建Soft-max Loss函数

其中是类别标签，网络不再单纯用one-hot向量计算Loss，而是使用类别的上下文（表现为label类别和其他类别的关系），直觉上这给了网络预测非label类别的更多可能（比如标签为iPhone，表示由iPhone拍摄，但拍摄内容和iPhone无关，那么我们倾向于网络预测图片的主体而不是iPhone）。用于表征label-flip噪音。为了避免语义偏移，本文中的关系图不更新。

#### 确定目标位置

现在我们训练出了CNN分类器，进一步我们需要清洗数据确定目标位置，以训练目标检测器。这个任务并不简单，因为：

1. CNN只能分离出一系列内部相似的分类，对视觉世界中负面的信息不敏感（比如背景噪声）。
2. 分类任务的Loss促使特征表示具有空间不变性。

这对确定目标任务很不利。作者提出了基于子分类探索的方法，参考Enriching Visual Knowledge Bases via Object Discovery and Segmentation。

**Seeds：**使用Google反馈的图片作为边框种子，这基于上文提到的，靠前的搜索结果往往具有目标居中且背景纯净的特点。

**Nearest Neighbor Propagation：**对每个种子，利用训练得到的fc7（全连接层7）特征训练E-LDA检测器。在所有下载图像上计算E-LDA的负统计数据，随后E-LDA在剩余图像中寻找前k邻近的box。为了效率，不遍历所有的box，使用EdgeBox来生成候选box。

**Clustering into Subcategories：**将同一类别的检测进行自底向上迭代聚合，行程基于E-LDA相似度得分和密度估计的子类别。这和NEIL的方法不同但取得了相似的结果且更具效率。

最后对每一个分类训练一个R-CNN检测器，负样本使用从YFCC数据集中随机选取的patch，正样本除了使用上文得到的正样本，作者还引入了两个数据增强手段：

**EdgeBox Augmentation (EA):**不断用EdgeBox寻找图像中感兴趣的区域，一旦出现和正样本具有的重叠，就把其加入训练集中。

**Category Expansion (CE):**利用关系图，在互联网上寻找更多相思类别的图像。在WorldNet上核验相似度后，将其加入训练集。

这些数据增强只用于帮助训练目标检测，CNN表示不会变化。

## 结论

作者提出了两步训练的方法使得网络可以接受噪声图像，通过简单图形训练CNN网络，随后利用此时的特征构建关系图，然后利用更为复杂但bias更小的数据，在关系图的约束下，fine-tuning网络，最终取得比较好的视觉表示。

作者提出需要区分webly supervised和unsupervised learning，基于网络的监督学习更适合语义任务比如检测、分类。无监督学习更适合generic task，如3D理解，grasping task（掌握任务）。