指导教师：_____ 提交时间：_____

The task of Digital Image Processing

数字图像处理

School of Computer Science

No: 1

姓名：马本腾

学号：2012302583

班号：10011202

基于网络监督的可视化概念学习

Santosh K. Divvala□,y, Ali Farhadiy, Carlos Guestrin

Figure 1

我们引入一个完全自动化的方法,任意给出一个概念,寻找得到一个详尽的词汇解释所有的外观变化(行动、交互、属性等等),并且训练出成熟的检测模型。这张图片显示了同一个物种的许多变化,我们的方法学会了四种不同类型的概念:对象(马),场景(厨房),事件(圣诞节),和行动（行走）。

摘要

对于识别的研究正从理论走向现实应用。虽然令人鼓舞的是看到它的潜力正不断被挖掘出来，但它同时也给那些做视觉领域研究的人们带来了一个基本的挑战:可扩展性。我们如何学习一个模型，对于任何概念,详尽涵盖所有的外观变化,而需要最小限度的或根本不用人力来收集视觉差异的词汇,收集图片和注释,学习这个模型呢?

在这篇文章中,我们提出一个完全自动化的方法，在给出的任意概念里，学习一个广泛模型的大范围的变化(如行为、交互、属性及其以后的状态)。我们的方法利用大量在线书籍的资源来寻找到那些描述视觉差异的词汇,并在数据采集和建模的步骤方面解放人力来训练模型。我们的方法在组织一个概念的视觉知识方面既方便又很有用,在视觉和 NLP 支持各种各样的应用。用户使用我们的在线系统时，查询过几个有趣的概念包括 早餐,甘地,美丽,等。到目前为止,我们的系统模型在 150 个概念中能涵盖超过 50000 变化,并为 1000 多万张图片在边界框加

了注释。

1.介绍

我们怎样才能了解一切(视觉)概念呢?这个问题主要有两个主轴。所有轴对应于所有可能出现的变化概念,而不同轴对应于不同范围的概念学习的可视化模型。学习一个概念模型的传统范式是:首先发现视觉空间概念差异(差异发现),然后收集训练数据即图片和注释,最后设计一个有力的模型,可以处理所有能发现的变化(差异建模)。对于差异发现,常见的做法是根据基准数据集[47]人工词汇。差异建模,这通常是独立于发现的步骤,大多数的方法使用一个分而治之的策略,把训练数据分成更小的子类[13]。有几种分类方法:视角[9],长宽比[18],poselets[5],视觉短语[43],分类标准[11]和属性[16,23]。

虽然上述范式有助于提前认识一些群体,但两个基本的和务实的问题仍未得到解答:首先,我们怎样才能确保一个概念所有方面的信息都被学习了?更具体地说,我们如何收集一个概念的详尽的词汇,涵盖所有的视觉差异?其次,我们如何规模以上范例学习一切?也就是说有可能设计出一种方法,减轻人力监督来发现词汇,收集训练数据和注释,能学习的模型?在本文中,我们引入一个叫做“网络监督”的方法来发现和建立视觉差异的模型。我们展示如何自动收集详尽的视觉差异,在没有人力监督的情况下学会使用可靠的视觉模型。

1.1 寻找差异并建模

几乎所有先前的工作都依靠直接人力监督来发现和建模。使用直接监督发现方法有几个缺点:

词汇不能穷举: 受文化、地理、现世性影响具有偏差。例如,“趟火墙”只有在世界的一些地方存在,因此可能被排除在“行走”的词汇之外。当抽样收集数据的视觉空间,有限的词汇量会导致高度偏差数据集[47]。词汇的穷尽性和模型的复杂性约束有一个权衡,词汇越详尽各个子类的差异就越小,从而可能减轻模型的复杂度。

特异性:预定义的词汇不能推广到新概念。例如,“饲养”可以鲜明的修饰“马”,但是不扩展到“羊”,而“剪切”适用于“羊”而不是“马”。这使得手动定义一个词汇的任务更加繁重。

灵活性:手工注释的词汇在创建数据集时会很死板(例如,属性列表[16,23],或视觉短语[43])。一旦注释被收入到数据集就很难在修改了,而这些数据最终决定用于处理数据的方法。例如,基于马的品种的分组(栗色的马,鞍马,等等)在Imagenet[11]不是很有用，,基于行为的分组(“跳马”,“控制马”,等)可能是更可取的。因此,如果注释能够基于概念的特征表示和学习算法被重新修饰就更好了。

可扩展性：当新的概念出现时要么产生新的数据集，要么往已有的数据集中添加新的注释.例如，在短语的情况下，可以让[12]和属性[16，23]，所有的 PASCAL VOC 的图像不得不添加新的注解。此外，如在建模步骤通常独立于发现步骤，用于模拟帧内概念差异得到的注释是经常不同，并且不相交的那些过程差异发现聚集。

1.2 概述

在这项工作中,我们提出一个新的方法来自动发现和建立视觉空间概念模型,避免了以上的限制(见图 2), ,我们利用大量资源网上的书(Google books ngram[33])发现差异词汇。这个发现词汇不仅广泛,而且 concept-specific。给定的一个术语。,‘马’,语料库包括 ngram 包含所有方面的术语,如行为(“喂马”),交互(“马装进桶里”)、属性(“停滞的马”)、视角(“马 前”),和其他(见图 1,顶部行)。

对于视觉差异的建模,我们建议关注的词汇发现和模型学习两个步骤。我们的方法不需要人力监督来给图片注释,从而提供更大的灵活性和可伸缩性。为此,我们在基于文本的网络图片搜索引擎,利用最近的进展和弱监督对象定位方法。过去的几年里,，图片搜索得到了极大改进。在现在可以检索有关集 对象—集中图像(感兴趣的对象占据了大部分的图像)的查询范围。虽然结果并不完美,对于大多数查询排名最高的图像往往是非常相关的[32]。最近成功的可变形的部分模型(DPM)，探测器[18],weakly-supervised 对象定位技术(36、39)已经重新引起人们关注。虽然这些方法不适合处理一组高度多样化的和受污染的图像,如图像检索‘马’，当面对一个相对干净的和以对象为中心的图像时，这些方法工作得非常

好,例如图像检索“跳跃的马”。

我们的相互交织的发现和建模的步骤的想法是部分地由该数据集的 VOC 通过下载图像使用一个明确的组查询扩展为每个对象编译观察而产生的激励(见表1[15])。

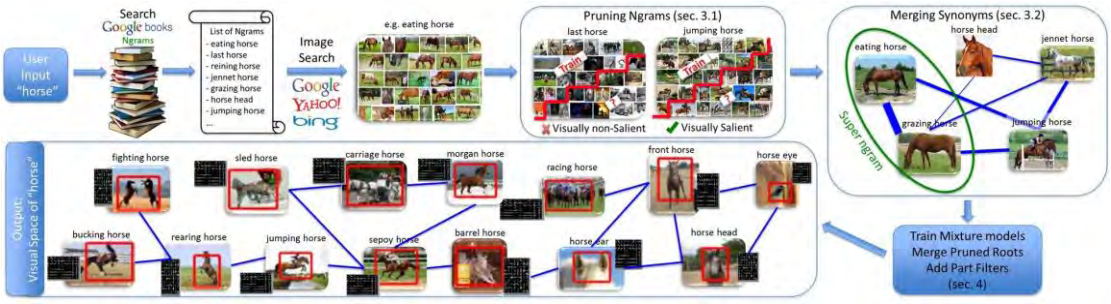

Figure 2: Approach Overview

然而,VOC 组织者检索图像后会丢弃关键字,他们可能认为关键字只对创建数据集有用,而对于模型学习并没有作用,又或者认为那些关键字是人工手选的有一定的局限性,过多的关注这些关键字并不能产生概括新的类别。在这项工作中,我们提出的了系统的查询扩展的想法减少了偏差数据的收集,并且能在无人监督的情况下学习更可靠的模型。

我们的成果包括:(1)对任意概念,没有明确人工监督的情况下,发现一个全面的词汇表(行为、交互、属性及其以后),和训练彻底检测模型,包括场景、事件、行动,地方,等。(2)实质性改进了现有的弱监督最先进的方法。对于某些类别,我们的结果与最先进的监管方法一致。(3)首次提供无人监管的发现活动的结果 (4)一个开源的在线系统,给出任何查询概念,自动学习一切视觉。迄今为止,我们的系统已学会了 50000 多个视觉系统模型,跨度超过 150 的概念,和注释超过 1000 万图像边界框。

2. 相关工作

改良内部类差异:之前约束内部类差异已经考虑基于长宽比的简单注释[18],[9],观点和特征空间聚类[13]。这些注释只能处理简单对象的外观变化[51]。近期作品已经考虑更复杂的注释,如短语[43],phraselets[12],和属性[16,23]。而明确

的监督需要收集短语和它们的边界框的列表[43], [12]就需要强大的监督来在数据集中注释所有对象的位置。虽然[24，27]直接从对象边界框就可以发现短语,但对象组成的短语词汇是有限的,并且不能发现复杂的动作,如,勒马,马把人摔下来,等。此外,所有的方法(12、24、27)发现短语只涉及数据集中完全注释的对象,即,他们不能发现“马 电车”或“桶马”当电车和桶没有注释。

属性[16,23]往往模棱两可的独立使用相应的对象,例如,一个“高个子”兔子比一匹“矮个子”马 还要矮;“cutting”指的是一个与马 相关的运动而对羊则有完全不同的意义。迄今为止,对于一个给定的数据集[37] 不存在一个描述他属性的清单。Weakly-supervised 对象定位: 由于最近 DPM 探测器[18]的成功，不使用边界框就从图片和视频训练检测模型的想法受到了重新关注[2、36、39,45]。虽然进展令人鼓舞,但有一些局限性有待解决。现有的基于图像的方法[2,36,45]当目标对象是高度混乱或当它只占一小部分的图像(如 bottle)效果并不好。视频的方法[39]依靠运动线索,从而不能本地化(如静态对象 tvmonitor) 。

最后,所有现有的方法在一个 weakly-labeled 数据集训练他们的模型,假设每个训练图像或视频包含的目标对象。在规模数以百万计的类别中,将这些方法改编一下，直接从嘈杂的 web 图片学习模型是更可取的。从 web 图片中学习:由于检测任务的复杂性和更高的监管要求,之前大多数的工作[38, 4,19,28,44,48]在使用 web 图像时只关注学习模型图像分类。(21,41)的工作重点是从一个大量的 web 图片库中发现常见的部分,但不报告定位结果。工作[49]使用主动学习的方法从 Turkers 边界框收集注释。[7]的工作旨在从 web 图像中发现常识知识,而我们的工作重点是学习详尽丰富的语义模型来捕获 intra-concept 差异。我们的方法会产生良好的模型,实现先进的性能基准 PASCAL VOC 数据集

3. 寻找差异性词汇

为了获得修饰一个概念的所有关键字,我们使用 Google books ngram English 2012 corpora [33]。我们专门使用 the dependency gram 数据,其中包含词类(POS)标记头=>修饰符对单词之间的依赖关系,并比原始 ngram 数据[30](见 4.3 节)更丰富。我们选择 ngram 数据而不是其他词汇数据库(比如 Wordnet 或维基百科列表[1]),因为它是更详尽,通用,包括流行(频率)的信息。books ngram 数据帮助我们使用覆盖所有纸面上的概念的变化。

给定一个概念及其相应的 POS 标记,如“阅读,动词”，我们就能找到词汇数据库所有有这个 POS 标记的注释。使用 POS 标记有助于部分消除歧义的上下文查询,如 reading action (verb) vs. reading city (noun)。the ngram dependencies 的所有检索给定概念,我们选择那些修饰符标记为名词,动词,形容词或副词[1]。我们总结多年的频率。使用这个过程,我们通常与大约 5000 ngram 概念。

不是所有使用上述收集过程的 ngram 视觉效果突出,例如 'particular horse', 'last horse',等等。面对这种干扰，我们的模型学习过程(第四节)仍然健壮,没必要再为不相关的 ngrams 训练一些成熟的探测器。为了避免不必要的计算,我们使用一个简单的和基于快速 image-classifier 的修剪方法。我们修剪步骤可以被视为一个级联策略的一部分,训练强模型之前拒绝弱模型.

3.1 分级修剪

这里的目标是从一个概念的 ngram 中确定视觉突出的 ngram。我们主要的直觉是视觉上突出 ngram 应表现出可预测的视觉模式访问标准的分类器。这意味着一个基于图像的分类器训练了一个视觉突出 ngram 应该准确地预测 ngram 看不见的样本。

我们首先为一组 ngram i 检索一组图片 I_i。为了保持低延迟,我们只使用缩略图(64×64 像素)的第一个 64 图像从图像搜索检索。我们都忽略 near-duplicate 图像。然后我们将这组随机分成大小相同的训练集和验证集 Ii = {Iit,Iiv},增强训练图像的镜像版本。我们还收集一个随机的背景图片 $\bar{I} = \{\bar{I}^t, \bar{I}^v\}$。对于每个 ngram,我们训练一个线性 SVM[6] W_i 作为积极训练图像I_i^t作为消极训练图像,使用密集的[HOG]特性[18]。这个分类器然后评估相结合的验证图像$\{I_i^v \cup \bar{I}^v\}$。

我们声明一个视觉上突出的 ngram i 如果平均分类器的精度(A.P.)[15]Wi 超过一个阈值。我们设置阈值较低的值(10%),以确保所有潜在的突出 ngram 传递到下一个阶段,只有完全无关的被丢弃。虽然我们的数据是嘈杂的(下载的图片没有手动验证是否包含研究对象)的概念,并且我们使用的 HOG+ linear SVM 框架并不是流行的图像分类,我们发现我们的方法能有效的和充分的修剪无关的 ngram。修剪步骤完成后,每个概念通常最终能得到约 1000 个 ngram。

3.2. 空间的视觉差异

在修剪 ngram 有几个同义词列表项，例如 ‘sledge horse’ 和 ‘sleigh horse’, ‘plow horse’ 和 ‘plough horse’, 此外,一些非同义 ngram 对应于视觉相似的实体, 如‘eating horse’ 和 ‘grazing horse’ (参见 Figure 2) [31]，为了避免相似词汇的重复训练，并从训练数据收集有用的数据，我们需要样品的视觉空间概念更仔细一些。我们如何能确定代表 ngram 跨概念的视觉空间?我们专注于两个主要标准: 品质和覆盖(多样性)。我们用 G={V,E}来代表 ngram 的空间，每个节点代表一个 ngram，每个边缘代表他们之间的视觉相似性。我们代表所有的空间 ngram 由图 G = { V,E },每个节点代表一个 ngram 和每个边缘代表它们之间的视觉相似性。

Cancer	{subglottic cancer, larynx cancer, laryngeal cancer} {rectum cancer,colorectal cancer,colon cancer}
Kitchen	{kitchen bin, kitchen garbage, kitchen wastebasket} {kitchen pantry, kitchen larder}
Gandhi	{gandhi mahatma, gandhi mohandas} {indira gandhi, mrs gandhi}
Christmas	{christmas cake, christmas pie, christmas pudding} {christmas crowd, christmas parade, christmas celebration}
Angry	{angry screaming, angry shouting} {angry protesters, angry mob, angry crowd}
Doctor	{doctor gown,doctor coat} {pretty doctor,women doctor,cute doctor} {examining doctor, discussing doctor, explaining doctor}
Apple	{apple crumble, apple crisp, apple pudding} {apple trees, apple plantation, apple garden} {apple half, apple slice, apple cut}
Jumping	{jumping group, jumping kids, jumping people} {jumping dancing, jumping cheering} {wave jumping, boat jumping}
Running	{running pursues, running defenders, running backs} {fitness running, exercise running} {running shirt, running top, running jacket}

Table 1: Examples of the vocabulary discovered and the relationships estimated for a few sample concepts.

每个节点都有一个数值 di 对应 ngram 分类器的品质。我们将数值 di 作为分类器的比重 Wi 的验证数据{ Iiv∪I¯v }。边缘权重 eij 对应 i,j 之间对应的视觉距离。测量的分数 jth ngram 分类器在第 i 个 Wj $(R:\mathbb{R}^{|\bar{I}^v \cup I_i^v|} \mapsto \mathbb{N}^{|I_i^v|})$ ngram 验证集{ Iiv∪I¯v }。为了避免问题未校准的分类器的数值,我们使用一个排序。排序函数将对 ngram 的背景图像库的验证集进行排序。符号 $R_{i,j}$ 对应的图像 I_i^v 和 $\bar{I}^v$ 之间用 W_j 计数。我们使用归一化平均排序作为边缘的比重[0 1]. $e_{i,j} = \frac{Median(R_{i,j})}{|I_j^v|}$ eij 的取值范围是[0,1]

寻找具有代表性的 ngrams 子集这个问题可以表述为，寻找子集 S ⊆ V，使得子集中的 F 最大：

$$\max_{\mathcal{S}} \mathcal{F}(\mathcal{S}), \quad \text{such that } |\mathcal{S}| \leq k, \tag{1}$$

$$\text{where } \mathcal{F}(\mathcal{S}) = \sum_{i \in V} d_i \cdot \mathcal{O}(i, \mathcal{S}). \tag{2}$$

O 是一个软覆盖函数,隐式地推动多样性：

$$\mathcal{O}(i, \mathcal{S}) = \begin{cases} 1 & i \in \mathcal{S} \\ 1 - \prod_{j \in \mathcal{S}} (1 - e_{i,j}) & i \notin S \end{cases} \tag{3}$$

这个方法搜索一个视觉可控的 ngram 子集 (有可靠 ngram 分类器)和覆盖概念里的的差异空间(类似于[3、22])。幸运的是,这一目标功能是一个子模块,因此在一个常数的近似最优解存在一个贪心算法。我们使用一个迭代的贪心算法,增加了在每个阶段 ngram i 提供最大增益在当前的子集(argmaxi F(S∪i)−F(S))。

Figure 3: 用我们的方法产生的样品定位结果，每行显示学到的 HOG 模板和一些训练实例。

这个算法提供了有代表性的 ngrams 的子集，能够在固定预算 k 的情况下最优

的描述空间差异。我们也可以使用相同算法合并相似的 ngrams 生成一个 superngrams。通过设置加入 S 中类似 n 元组到真正高增值的代价，每个 NGRAM L /∈S 能合并到 S 最接近的成员。

我们的合并过程通过合并视觉类似的行动,交互,和属性揭示了有趣的 ngram 关系。例如,我们的方法发现以下 ngram“horse”的视觉相似。{tang horse, dynasty horse}, {loping horse, cantering horse}, {betting horse, racing horse}, etc。表 1 展示了更多其他概念的的例子。使用这种合并方法，我们将 Ngram 的数量减少到大约 250 个 supergram。

4. 模型学习

训练的图像探测器收集使用图片搜索与查询短语作为 gram 构成 supergram. 我们为每个查询图像类型下载 200 个全尺寸,完整的颜色,“照片”。我们调整图片最多 500 像素(保持长宽比),丢弃所有 near-duplicates,忽略图像与极端的纵横比(长宽比> 2.5 或< 0.4)。我们下载的图片分成训练集和验证集。训练的根源:Pandey et al .,[36]证明 DPM 探测器使用弱监督[18]的可能行。直接将他们的方法应用到一个概念中所有的图像(汇集所有 ngram)结果得到一个不太好的模型[1]。因此我们为每个存在视觉差异约束的地方单独训练一个 DPM ngram。(36)初始化数据处理机与完整的图像的边界框。我们发现使用这个初始化通常会导致图像边界的边界框陷在潜在重新集群步骤[1]。为了规避这一问题,我们初始化边界框内的子图像图像,忽略了图像的边界。使用这个初始化还避免了两级训练过程中使用[36],在第一阶段潜在的根位置识别和剪裁,DPM 在第二阶段训练。

类似于[18],[36]也使用宽高比启发式初始化组件。这是次优弱监督设置，因为图像长宽比是一个不太好的启发式聚类对象的实例。为了解决这个限制,我们使用特征空间聚类初始化模型提出了[13]。而我们 ngram 词汇有助于隔离一个概念的主要外观变化,下载的图片每 superngram 还有些剩余的外观变化。例如,“跳马”ngram 马跳的图像在不同的方向。应对这样的外观变化,我们使用组件初始化特征空间聚类的混合物。在嘈杂的 web 图像,这个过程提供了一个健壮的初始化。一些混合组件作为噪声下沉,从而允许干净的模型学习[51]。通常在我们的实验中,我们发现,每个 ngram 的 70%的组件作为噪声下沉。训练这种带有干扰的模型是很浪费的。因此,我们首先训练根过滤器，为每个组件修剪嘈杂的成分。

修剪嘈杂的组件:修剪嘈杂的组件,每个组件运行探测器的验证集和评价其性能。考虑到积极的实例验证集内的每个 ngram 真实的边框和组件标签,作为一个潜在的图像分类问题来处理。具体地说,我们第一次运行 ngram 组件混合，探测器在其完整的验证集(伸出正面形象以及随机的背景图像)。然后我们记录每个图像的最高检测和使用的组件标签检测隔离图像。我们现在有一个隔离池验证每个 ngram 的图像的组件。没有真实的盒子,我们假设前检测的积极图像是真实的而消极的图片是错误的,因此计算的平均精度(A.P.)仅使用检测成绩(忽略重叠)。我们声明一个组件如果它 A.P.低于一个阈值(10%)或者其训练或验证数据太少(<5)的积极的实例那么它就是嘈杂的。第二个条件有助于我们丢弃不符合范例的附带图片的组件。而根滤波器模型相比相对较弱的部分模型,我们发现他在修剪嘈杂组件方面的工作还是很有效的。

合并修剪组件:一些组件在不同 ngram 探测器最终学习相同的视觉概念。例如,“hunter horse”实例的一个子集,非常类似于“jumping horse”实例的一个子集。合并步骤 3.2 节中被认为是单一的分类器训练与完整的图像特征。组件模型的混合更精确 (本地化实例使用检测窗口),他们可以识别微妙的相似之处,这些差异从整个图像角度出发很难被发现。我们遵循一个类似的过程如 3.2 节中列出，选择一个代表性的子集组件和类似的合并。\

具体来说,我们代表所有 ngram 组件的空间由一个图 G =(V,E },其中每个节点代表一个组件,每个边代表它们之间的视觉相似性。现在的数值为每个节点 di 对应于组件的质量。我们将它设置为 AP 的组件(在上述修剪步骤计算)。每边的比重 ei,j 同样定义为中位数排名 jth 组件运行探测器获得的第 i 个组件验证集。(我们继续使用每个图像的顶部检测得分,假设检测积极图片是真,消极图片是假)。我们解决相同的目标函数与方程(1)中列出的选择具有代表性的子集。我们发现这个子集选择步骤导致约 50%更少的组件。组件的最终数字平均约 250 /概念。图 3 显示了我们的一些发现组件。

考虑到代表组件的子集,我们终于增强他们部分中描述的[18],并随后合并所有的组件来生成最终的探测器。

5. 结果

我们提出的方法是，可用于训练广泛检测模型为任何概念的通用框架。为了定量地评价我们的方法的性能，我们提出业绩目标和行动 detection.Object 检测：我们评估了我们训练的检测模型的表现为 20 个教学班，在 PASCAL VOC2007 testset[15]。我们选择这个数据集作为最近的最先进的国家监督弱的方法已被评估它。在我们的评测中，我们保证没有任何的在 2007 年 VOC 的 testset 测试图像存在于我们的训练中。

表 2 显示的结果获得通过我们的算法和比较一个国家的最先进的基线[39，45]。[45]采用弱的人的监督（与映像级 VOC 数据对于训练标记物），并从对象性初始化[2]即依次经过培训上的 VOC2007 子集 4。相比之下，我们的方法是使用网络的监管，因为即使没有图像提供培训 5。尽管如此，我们的结果大大超过了以前的最好成绩弱监督检测对象。

图 1 示出了一些 行动，相互作用，并学会了“马”的属性。图 4 显示了获悉，为其他概念的车型。

动作检测：将 VOC 的挑战[15]被假定的边界框为人类执行的动作被称为动作分类任务（无论是在测试和列车图像）和人类活动具有被识别。我们考虑操作的检测，这里不仅确实在图像的动作必须被识别，而且还有局部与边界框的更有挑战性的任务。此外，我们尝试执行动作的检测中无监督的设置，甚至在训练图像是不能提供的。据我们所知，这是在网络监管动作检测的 VOC 数据集的第一次尝试。对于强监督动作检测的 VOC2011 VAL 组获得[12]报道的结果最近的工作。表 3 报道了结果，并将其与报道[12]的监督基线。

图 1,4 显示一些模型了解到的 VOC 动作类’跳’，’跑’，和’走’。对于每一个 Action 类，我们的做法学会了详细

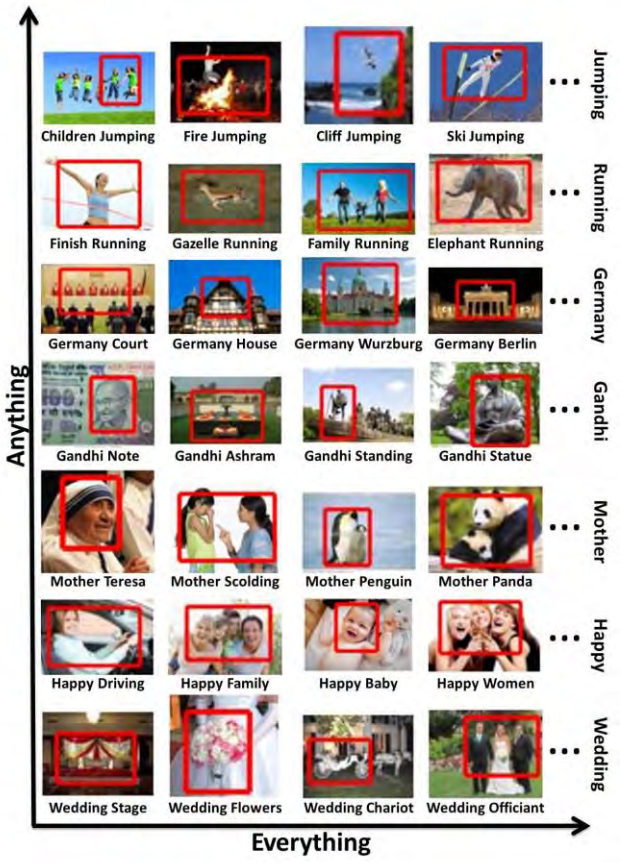

Figure 4: 对于任何的概念（y 轴）我们的方法可以学习大量的模型，每行显示四个的变化，我们的方法已经学会了。全部结果（详见词汇和训练的机型），可在我们的项目网站。

词汇量，揭示了几个细粒度的变化，例如，“步行模式”与“竞走”。类似于字典或百科全书提供了一个概念的不同词汇的内涵，我们的方法产生了不同的视觉内涵。我们也跑了我们的实验在复合“骑自行车的类别，但发现我们的模型表现不佳（AP 为 4.5%和 41.6%[18]）作为 VOC 地面实况中只涵盖执行操作的人，而我们的无监督的方法也本地化
随着人的自行车。 我们的方法得到的 31.6%的 AP 时的重叠标准与地面的真理降低到 25%）。

Method	Supervision	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbik	pers	plant	sheep	sofa	train	tv
[45]	weak	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0
[39]	weak	17.4	-	9.3	9.2	-	-	35.7	9.4	-	9.7	-	3.3	16.2	27.3	-	-	-	-	15.0	-
Ours	web	14.0	36.2	12.5	10.3	9.2	35.0	35.9	8.4	10.0	17.5	6.5	12.9	30.6	27.5	6.0	1.5	18.8	10.3	23.5	16.4

Table 2: 结果（AP）上 VOC2007（测试）检测对象。行 1 和国家的最先进的 2 显示结果为弱监督检测对象。[45]与映像级标签 VOC2007 训练数据串并使用对象性进行初始化。[39]列车手动选择视频，但没有对 10/20 类（忽略类没有运动）边界框，并显示结果。第 3 行显示了我们的网络监督的结果，也就是说，即使提供的训练图像。第 4 行显示当前状态的最先进的成果的充分监督检测对象是一个可能的上限弱监督方法。我们的方法优于监督 DPM 对鸟类和狗，对羊基本和 DPM 一致。

	jumping	phoning	walking	takingphoto
[18] (supervised)	*6.1*	*4.1*	*10.9*	*1.1*
Ours	12.8	3.6	10.7	0.2

Table 3: 结果（A.P.）上 VOC2011（VAL）动作检测。上面一行显示的结果完全监督作用检测得到使用[18]一个国家的最先进的（如文献[12]），而下排显示我们的结果。评价协议是相同的物体检测（重叠>50％是成功的。）关于“跳跃”我们的成绩击败了全面监督结果，其余 3 类几乎持平。

Figure 5: 我们的方法学模型的概念，所有的视觉内涵。例如，名词“训练”既可以指一件衣服或机车，而名词“椅子”既可以指一件家具或指定。今后的工作中可以把重点放在探索按我们的方法发现的多义性。

我们发现几个问题：

重叠的程度：我们的最终检测模型是一种多组分模型（组件≥250 平均数）。给定一个测试图像，可能有几个有效的检测通过我们的模型，如马车的图像不仅有’曲线的马’的检测也是’马车’和’马头’检测。如在 VOC 标准要求为每个测试实例有 50％的重叠一个唯一的检测中，所有的其他有效检测被宣布为假积极或者是由于本地化不佳或多次检测。选择从有效检测池的正确框完全无监督的设置是一个具有挑战性的研究课题。

一词多义：我们的框架学习的概念，比如一个通用模型，汽车模型包括一些公交样车部件，而 VOC 数据集专门侧重于典型的汽车（而且，从公共汽车歧视车）。这样的多义性是一个致命的缺陷，在与词汇资源处理作为相同术语可以指两个完全不同的概念（参见图 5）。为了缓解这些问题，有可能调整我们的模型来解释数据集的偏差，从而提高其性能[20，47]。在完全无监督设置调谐偏差也一个有趣的研究方向。

测试我们的模型涉及卷积每个测试图像的大约 250 组件的详尽的模型。这个测试步骤可以通过利用最近的快速检测方法[10，14，46]容易地加快。

6. 结论与应用前景

我们提出了一个完全自动化的方式来显示 覆盖任何概念的详细的词汇，培养全成熟的检测模型为其。我们已显示几个概念的结果（包括物体，场景，事件，行为 和地点）在本文中，通过使用我们的在线系统可以得到多个概念我们的方法使一些 未来的应用和研究方向成为可能：

指代消解：NLP 中的一个核心问题是要确定当两个文本提到的名字是同一个实体。这里最大的挑战是无法推导语义知识。例如，斯坦福大学最先进的艺术国度系统[25] 在下面的句子：英迪拉·甘地是印度第三个首相。圣雄甘地是印度民族主义领袖。甘地夫人

灵感来自于圣雄甘地的著作。没有链接到“圣雄甘地“和”太太甘地“到”英迪拉·甘地。我们的方法能够链接到圣雄甘地圣雄甘地和英迪拉·甘地夫人甘地（见表 1）。我们设想，用我们的方法所提供的信息应为指代消解有用的语义知识。

释义：重写文本短语换句话说，同时保留它的语义是一个活跃的研究领域 NLP。我们的方法可以用于发现复述。例如，我们发现一个“grazing horse ”在语义上非常类似于一个“eating horse ”。我们的方法可以用来产生一个语义相似性得分为文本短语。

时间演化的概念:有可能存在模型沿时间轴概念的视觉差异。我们可以使用以年为频率信息

使用 ngram 语料库来识别一段时间,然后学习模型(参见图 6)。这不仅可以帮助学习概念的演变[26],而且可以自动检测实例[35]。

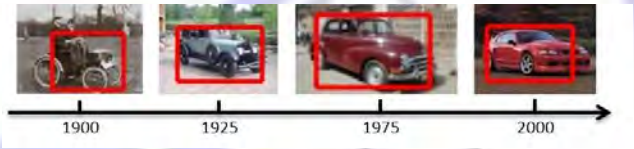

Figure 6: 我们的方法可以用来学习一概念的时间演化。该图显示了自 1900 年以来我们的合并算法合并的模式，1950 年与 1925 年和 1975 年（可能表明汽车的外观在这段时间内没有发生重大变化的概念“车”与训练每一代独立的组件（25 年）的实例)。

更深层次的图像解释：最近的作品都强调提供对象检测更深的解释，而不是简单地与边界框标注它们[34，43]的重要性。我们的工作由生产提高了检测的任何概念证实了这方面的研究。例如，除了对象边界框（如，“马”），它可以提供

对象的一部分盒（例如，“马头”，“马足”等），也可以注释对象操作（例如，’战斗’）或对象类型（例如，’驴骡马’），由于我们使用对应于真实世界实体的 ngram 标签，但也可以以检测直接链接到其相应的维基百科页面来推断详情[42]。

了解操作：操作和互动（例如，“斗马”，“驾驭马”）过于复杂，用简单的原语来解释。我们的方法有助于发现了一个全面的词汇，涵盖的任何行动都（细微的）差别。例如，我们已经发现了行走动作超过 150 种不同的变化包括“球走’，’情侣散步”，“框架行走”（见图 1，下排）。这样一个详尽的词汇有助于生成图像详细的描述[17，29，34，40，50]

分割及发现：在我们的模型中的每个组件具有的所有紧密的外观空间（参见图 3）对准的训练实例。因此，它是可以 cosegment 的实例，以及使用 cosegmentation[8，21，41]学习一个前景分割模型的每个组件。这使我们的做法延伸到执行无监督像素级分割，并对任何的概念取得了丰富的语义细分模型。

致谢：这项工作是由 ONR 支持 N00014-13-1-0720，ONR 穆里 1141221-293180，ONR PECASE N00014-13-1-0023，NSF IIS-1218683，和 NSF IIS-1258741。我们对库马尔的建议表示感谢。

参考文献：

[1] Wikipedia list of lists of lists. http://en.wikipedia.org/wiki/List of lists of lists.
3
[2] B. Alexe, T. Deselares, and V. Ferrari. Measuring the objectness of image
windows. In PAMI, 2013. 3, 6
[3] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse
M-best solutions in markov random fields. In ECCV, 2012. 4
[4] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve
object classification: a domain adaptation approach. In NIPS, 2010. 3

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human
pose annotations. In ICCV, 2009. 1
[6] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines.
2001. 4
[7] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge
from web data. In ICCV, 2013. 3
[8] X. Chen, A. Shrivastava, and A. Gupta. Enriching Visual Knowledge Bases via
Object Discovery and Segmentation. In CVPR, 2014. 8
[9] O. Chum and A. Zisserman. An exemplar model for learning object classes. In
CVPR, 2007. 1, 3
[10] T. Dean et al. Fast, accurate detection of 100,000 object classes on a single
machine. In CVPR, 2013. 7
[11] J. Deng et al. Imagenet: A large-scale hierarchical image database. In CVPR,
2009. 1, 2
[12] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational
phraselets. In ECCV, 2012. 2, 3, 6, 7
[13] S. K. Divvala, A. A. Efros, and M. Hebert. How important are ‘deformable
parts’ in the deformable parts model? In ECCV Workshop on Parts and Attributes,
2012. arXiv:1206.3714. 1, 3, 5
[14] C. Dubout and F. Fleuret. Exact acceleration of linear object detectors. In
ECCV, 2012. 7
[15] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The
pascal visual object classes (voc) challenge. In IJCV, 2010. 2, 4, 6
[16] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their
attributes. In CVPR, 2009. 1, 2, 3
[17] A. Farhadi et al. Every picture tells a story: Generating sentences for images.
In ECCV, 2010. 8
[18] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object
detection with discriminatively trained part based models. PAMI, 2010.
http://www.cs.berkeley.edu/ rbg/latent. 1, 2, 3, 4, 5, 6, 7

[19] R. Fergus, F.-F. L., P. Perona, and A. Zisserman. Learning object categories
from internet image searches. In Proc. of IEEE, 2010. 3
[20] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the
damage of dataset bias. In ECCV, 2012. 7
[21] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In CVPR, 2012. 3, 8
[22] A. Kulesza and B. Taskar. k-DPPs: Fixed-size determinantal point processes.
In ICML, 2011. 4
[23] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object
classes by between-class attribute transfer. In CVPR, 2009. 1, 2, 3
[24] T. Lan, M. Raptis, L. Sigal, and G. Mori. From subcategories to visual composites:
A multi-level framework for object detection. In ICCV, 2013. 3
[25] H. Lee et al. Deterministic coreference resolution based on entity-centric,
precision-ranked rules. In Computational Linguistics, 2013. 7
[26] Y. J. Lee, A. A. Efros, and M. Hebert. Style-aware mid-level representation for
discovering visual connections in space and time. In ICCV, 2013. 7
[27] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for
scene understanding. In CVPR, 2012. 3
[28] L.-J. Li and L. Fei-Fei. Optimol: Automatic online picture collection via incremental
model learning. In IJCV, 2010. 3
[29] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image
descriptions using web-scale n-grams. In Computational Natural Language Learning, 2011. 8
[30] Y. Lin et al. Syntactic annotations for the google books ngram corpus. In ACL,
2012. http://books.google.com/ngrams/datasets. 3
[31] T. Malisiewicz and A. A. Efros. Recognition by association via learning perexemplar
distances. In CVPR, 2008. 4
[32] E. Mezuman and Y. Weiss. Learning about canonical views from internet image
collections. In NIPS, 2012. 2
[33] J.-B. Michel et al. Quantitative analysis of culture using millions of digitized

books. In Science, 2010. 2, 3
[34] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. From large scale
image categorization to entry-level categories. In ICCV, 2013. 8
[35] F. Palermo, J. Hays, and A. Efros. Dating historical color images. In ECCV,
2012. 7
[36] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object
localization with deformable part-based models. In ICCV, 2011. 2, 3, 5
[37] D. Parikh, A. Farhadi, K. Grauman, T. Berg, and A. Gupta. Attributes. In CVPR
Tutorial, 2013. https://filebox.ece.vt.edu/ parikh/attributes/. 3
[38] O. Parkhi, A. Vedaldi, and A. Zisserman. On-the-fly specific person retrieval.
In Image Analysis for Multimedia Interactive Services, 2012. 3
[39] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class
detectors from weakly annotated video. In CVPR, 2012. 2, 3, 6, 7
[40] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps
where and why semantic relatedness for knowledge transfer. In CVPR, 2010. 8
[41] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery
and segmentation in internet images. In CVPR, 2013. 3, 8
[42] B. C. Russell et al. 3d wikipedia: Using online text to automatically label and
navigate reconstructed geometry. In Siggraph Asia, 2013. 8
[43] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In CVPR, 2011.
1, 2, 3, 8
[44] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from
the web. In PAMI, 2011. 3
[45] P. Siva and T. Xiang. Weakly supervised object detector learning with model
drift detection. In ICV, 2011. 3, 6, 7
[46] H. O. Song, R. Girshick, and T. Darrell. Discriminatively activated sparselets.
In ICML, 2013. 7
[47] A. Torralba and A. Efros. Unbiased look at dataset bias. In CVPR, 2011. 1, 2,
7

[48] D. Tsai et al. Large-scale image annotation using visual synset. In ICCV, 2011.
3
[49] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training
object detectors with crawled data and crowds. In CVPR, 2011. 3
[50] L. Wu, R. Jin, and A. K. Jain. Tag completion for image retrieval. In PAMI,
2013. 8
[51] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes. Do we need more training
data or better models for object detection? In BMVC, 2012. 3, 5