指导教师：杨涛　　提交时间：2016/3/17

CVPR2015 Paper Translation

No：01

姓名：韩仁杰

学号：2013300216

班号：10011301

利用位置环境来改进图像分类

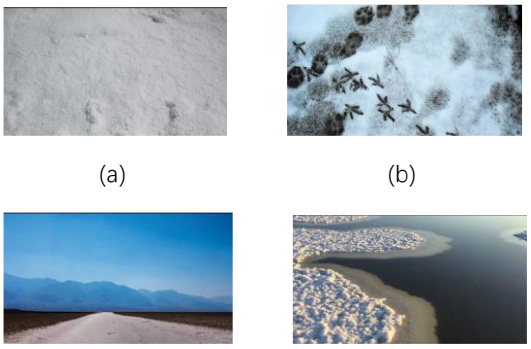
Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, Lubomir Bourdev
Computer Science Department, Stanford University Facebook AI Research
{kdtang,feifeili}@cs.stanford.edu {mano,robfergus,lubomir}@fb.com

摘要

随着带有 GPS 功能的移动电话和摄像头的广泛普及，现在上传到网络上的图片已经很普遍的都带有一些相关的 GPS 位置信息。此外，一些试图通过视觉特征来预测 GPS 位置信息的研究也让一些和 GPS 位置信息相关的问题暴露出来。在本文中，通过同时给出训练组和测试组的图片的 GPS 坐标吗，我们抓住了利用位置环境来进行图像分类的问题。我们采用了不同的方式来对 GPS 坐标进行编码和特征提取，并展示了如何将这些特征体现成为为目前最高水平的图像分类和问题识别技术——卷积神经网络。我们同时也给出了如何有可能在卷积神经网络框架的内部同时学习特征子集的最佳池半径。为了测试我们的模型并帮助提升在本领域的研究，我们使用了一系列位置敏感的概念并且注释了一个雅虎 Flickr Creative Commons 100M 数据集，这个数据集带有已经公知的与概念相关的 GPS 坐标信息。通过改变位置信息，我们可以在平均准确率均值的基础上达到 7%的增益。

1. 简介

正如图一所示，有时候如果没有给出环境信息，即使是人也很难分辨



图一：这些图片中哪一个是有雪？仅仅通过观察图片，可能很难说。然而，如果我们知道(a)拍摄于犹他大学的波利维尔盐洼，(c)和(d)则是在加利福尼亚州从来没见过雪的死亡谷和帕罗奥图，(d)是在暴风雪经常来临的新罕布什尔呢？图像信息需要补充的内容才能被确认。

出图片的内容。仅仅通过对图片的观察，我们可以得出结论这些例子都是关于雪的。然而，(a)的地点是在犹他大学的波利维尔盐洼，(c)和(d)则是在加利福尼亚州从来没见过雪的死亡谷和帕罗奥图，(d)是在暴风雪经常来临的新罕布什尔。通过这个信息，很容易就正确的得出结论，只有(b)是唯一的包含雪的图片。

出于这一观察的动机，我们抓住了利用位置环境来改进图像分类的问题。特别的，我们对用经常出现在网络上的概念对用户图片进行分类的方法很感兴趣。这些概念包括有人们经常去拍摄的物体，场景以及一些特定的地标，而网络也是最大的带有地理位置信息图片的来源。基于在 CNN 上

发表的[26]中介绍的构架，被认为是最具影响力的图像分类和识别提出了如何表现并将位置特征整合到网络构架中。这不是一个容易的为题，我们发现一些天真的想法，比如将 GPS 坐标与分类器连接在一起或者利用附近的图像作为一个先验的贝叶斯导致实验几乎没有收获。然而，通过知道 GPS 坐标，我们得以利用地理数据集和被各种各样的机构，旅行社收集的调查结果。我们在数据驱动方式上还可以利用大量被标注有 GPS 坐标的网络数据。

总而言之，这篇论文的贡献可以被归为三部分。

从全球定位系统中构建有效的位置特征。我们在已经给出的经度和维度坐标的基础上，拓展性的提出了 5 个不同类型的特征，并且对每个特征都进行了影响力的综合评价。

结合位置特征的网络体系结构。我们展示了如何将这些额外的特征添加进 CNN[26]。这让我们得以在一个共同的框架下根据不同的特征类型之间的联系学习它们的可视特征。除此之外，我们还给出了我们如何能在同一个框架下同时学习构建特征子集所必需的参数，使得我们能够改进性能并更好的理解什么是网络学习。

YFCC100M-GEO100 数据集。我们使用了雅虎 Flickr Creative Commons 100M 数据集[4]的一个子集来注解一系列位置敏感的概念，我们称其为 YFCC100M-GEO100 数据集，并将它的注解公开了。这一个数据集包含超过 100 种的 88,986 张图片，并且允许我们大规模的测试我们我的模型。

2. 相关工作

有很多研究都将焦点聚集在图片的地理定位问题上，比如确定静态摄像头位置[22]，城市规模定位识别[37]，利用图像的 GPS 定位[20,43]，位置识别[19,40]，路标识别[10,13,34,44]，利用地理信息确定地理位置[10,21,31,36]，还有基于图像表示的地理位置[12]。还有更多的工作尝试补充对应的图像数据[7,8,27,30,32]，例如数字高程图和土地覆盖调查数据，我们从中汲取灵感，进而构造我们的特征，和这些作品相比我们，我们假设我们在进行图像分类时已经有了 GPS 坐标。

除此之外，还有一些探索了图像和位置信息其他方面的工作，例如 3D 汽车的位置[33]，带有地理标记的图片整理[14]，网络照片上得结构[38]，辨认城市身份[45]，看到可见图像的背后[25]，发掘具有代表性的地理视觉元素[16,28]，通过图像预测土地覆盖[29]，使用典型相关分析的注释增强[11]。

最为相近的文章都是利用位置信息来识别任务[5,6,9,23,42]。在小型城市环境中利用移动设备的地理服务对处理对像的识别的文章[5]。在以将交通标志，交通信号，垃圾桶，消防栓和街道灯的精确位置信息投射在图

像上为前提的条件下，利用地理信息系统数据集[6]。利用鸟的踪迹来估计一个时空先验模型来帮助提高细粒度分类性能[9]。利用大量的地理信息来代表当地居民的信息[23]。利用季节和位置上下文的概率框架来帮助提高图像的地域识别[42]。我们的工作所不同的地方在于我们所感兴趣的是在互联网上除了鸟[9]，城市物体的集合[5,6]，简单事件[23]或是通用地域类型[42]的一些用于识别的概念集，并且构建一些不是特定于特定类别或来源的地理信息系统信息的特征。除此之外，我们详尽地探讨将这些功能集成到一个美国有线电视新闻网的方式，并且提出了一个将地理特征参数化的方式，以及允许网络学习最具区别的地理特征的扩展反向传播算法。我们还引入了从现实世界图像采集的大规模的地理标签数据集来训练我们的模型并有效地评价性能。

与之有密切关系的是从众多作品的上下文中，已证明是有助于计算机视觉下的各种任务[15,41]。我们通过考虑我们的图像的 GPS 坐标和提取互补的定位功能充分的利用了上下文信息。

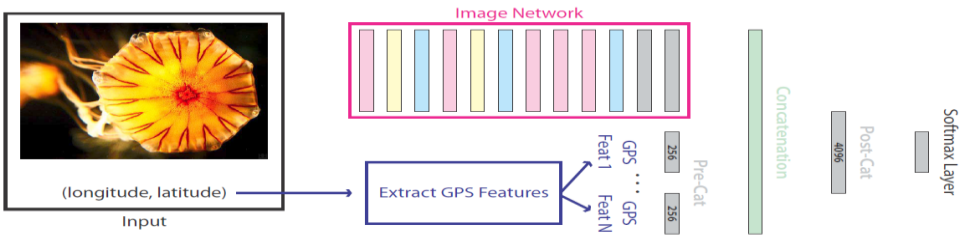
3. 我们的方法

和标准图像分类问题类似的，我们被给了一组带有相关分类标签 $\{y_1, y_2, \ldots, y_n\}$，的 n 个训练图像 $\{I_1, I_2, \ldots, I_n\}$，而 y 是我们所试图预测的类别集并且 $y \in C$。除了图像之外，我们还有每个图像的 GPS 坐标信息 $\{(long_1, lat_1), (long_2, lat_2), \ldots, (long_n, lat_n)\}$，$long_i$ 和 lat_i 则分别是图像 i 的经度和纬度。要注意的是我们在训练组和测试组都给出了 GPS 坐标，我们的目标是预测图像和 GPS 坐标两者的类标签。在本文中，我们侧重于美国本土的图像，但我们的绝大多数特征，都可以扩展应用于包括整个世界。

3.1. 神经网络结构

我们使用[26]中介绍的 CNN 模型作为基础，因为在图像分类识别中，这个模型和它的扩展是常用的基准[18,35,39]。至于要了解在网络架构方面的更多详细信息，我们建议读者阅读[26]。为了将位置特征合并进整个网络，我们在 softmax 层之前添加了一个层来连接不同的特征类型，正如图 2 所示。这使得作为旨在学习有效的图像过滤器和功能的 CNN 模型的较低层，有了更加直观的感觉，我们感兴趣的是稍后在一个更高的语义层次合并我们的特征。除此之外，我们还实验了在级联层之前和之后添加完全连接层来增加额外的深度，例如在图 2 中的 pre-cat 层和 post-cat 层，我们会在本文中的后面进行详细的综合性实验。

利用这一结构，我们面临着的是如何提取图像特征的问题。对于每一



图二：我们的 CNN 构架。粉红色的矩形表示卷积层，黄色的矩形表示归一化层，蓝色的矩形表示池层，灰色的矩形表示完全连接层，还有绿色矩形表示级联层。最终的完全连接层是 softmax 层。我们的模型输入时一个图像以及它相关的经纬度坐标。由品红框表示的图像网络是[26]中介绍的网络架构。

个图像 i，我们构造了一组能有效地代表上下文信息中关于由 GPS 坐标指定的位置的特征 $(long_i, lat_i)$。为了做到这一点，我们利用地理财富的数据集以及通过各机构收集的调查，其中记录了种类繁多的位置统计，调查覆盖的种类多样，从年龄和学历到地理特征，如高程和沉淀。我们还利用了大量的互联网提供的带有地理标记的数据，如图像和文本的帖子。

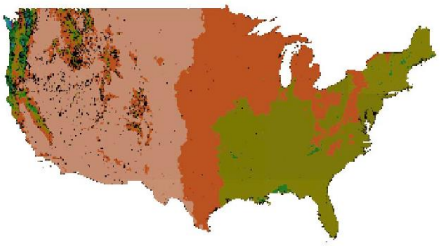
3.2. GPS 编码特点

实际的 GPS 坐标是非常精确的位置信息的指标，因此很难为分类器有效地使用。为了更好地利用坐标，我们将美国邻接以纬度比经度比例为 1 比 2 的比例构建一个矩形网格，为每一个图像构建一个指标矢量，这个矢量会指示哪些网格单元的 GPS 坐标图像 $(long_i, lat_i)$ 落入，产生的特征矢量的尺寸等于在网格单元的总数。被选择的纵横比，使每个网格单元是大致的正方形。我们使用高达 100×200 的矩形网格，得到 25x25km 正方形单元，受限制于计算时间和内存更大的网格。

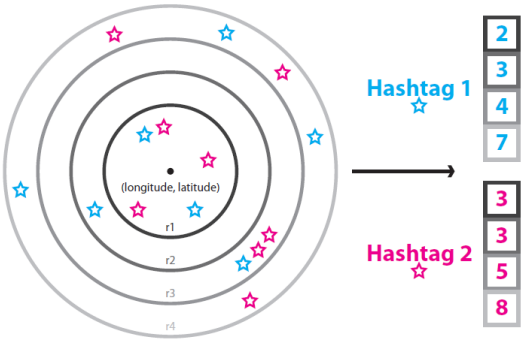
3.3. 地理图特征

存在着许多不同类型的地理地图和数据集，它们为用不同的颜色代表不同的地理特征的彩色地图提供每个 GPS 的详细信息坐标。尤其的，谷歌地图[2] 是许多存储大量的这种地图集的在线网站之一，用图 3 作为例子，我们使用谷歌地图的 10 种不同的地图：平均植被，国会选区，生态区，海拔，危险废物，土地覆盖，降水，太阳能资源，总能量和风力资源。由于每个地图使用不同的颜色以表示一个特定位置特征的值，对于每一个图像 i，我们对每个地图类型下点的 GPS 坐标 $(long_i, lat_i)$ 周围的像素颜色值都采取 17x17 补丁的方式进行标准化，并且将它们连接起来构成一个 8670 维的特征。直观地看，地图中的特征如降雨会告诉我们有多大的可能会看到

图片中的一把伞，而一些技术指标如高程的值可能会告诉我们有多大的可能是我们看到了雪。



图三：例如美国地图的地理降雨量[2]，有较深的颜色大致表明更大的平均降水量。降雨量多的地区可能会引起图像中更经常的出现一些对象如伞。



图四：为了构建标签背景特征，我们通过寻找标记有相关的 hashtag 标签和 GPS 坐标的 Instagram 图片，观察了围绕每个 GPS 坐标的 hashtag 标签的分布。对于每一个 hashtag 标签，我们通过计数出现一个特定的半径内的每个 hashtag 标签（蓝色的/洋红色的星星）的次数，划分出了不同半径的圆。

3.4. ACS 特征

在被给予了每个图像 i 的 GPS 坐标 $(long_i, lat_i)$ 之后，我们就可以通过执行反向地理编码，以获得相应的邮政编码。这使我们能够利用由邮政编码分类的由像美国政府这样的机构收集的丰富来源的地理调查。我们使用了美国社区调查（ACS）[1]，它是一个正在进行的调查，每年提供的数据会带有统计数据：年龄，性别，种族，家庭/人际关系，收入和福利，医疗保险，教育，退伍军人身份，残障，工作状态和生活条件下，所有这些数据都是通过邮政编码分类，汇集时间超过 5 年。我们把每个统计数据都作为一个特征，并综合它们形成一个向量，结果出现了一个 21038 维的特征向量。直观地说，一些统计信息，如年龄可能会告诉我们有多么可能会在图片中看到玩具，而统计数据，如收入可能告诉我们有多么可能在图像中看到昂贵的汽车。

3.5. 标签背景特征

上述地理地图和 ACS 特征都是基于各机构收集的关于一个特定的位置的地图和调查数据。然而，大量的数据直接位于互联网，它上面每天上载数以百万计的图像，其中很多标有 GPS 坐标。我们提出了一套数据驱动的特征来利用 Instagram[3]上的图片。

直观地，对于每个与 GPS 坐标 $(long_i, lat_i)$ 相关联的图像 i，我们的目标是捕获在附近的 hashtag 标签的分布。图像 i 周围经常出现的 hashtag 标签可以帮助指示在现实世界的情况下发生在图像 i 的周围的事情的类型，给我们关于在图像中的物体的上下文信息。我们首先来定义一组我们感兴趣的 hashtag 标签，记为 H。对于一个特定的 hashtag 标签 $h \in H$，我们将从

Instagram 得到的图像和 GPS 坐标与 hashtag 标签匹配。然后，我们定义一组半径，记为 R，对于每个半径 $r \in R$，我们在 $(long_i, lat_i)$ 周围划分出一个半径为 r 的圆，并且计数带有 hashtag 标签 h 的图像落在该圆内的个数。正如图 4 中所展示的，这是以每个 R 中的半径委员，并且每个 H 中 hashtag 标签的计数完成，产生的一组 $|H| \times |R|$ 数值。

为了从这些数值中构建特征，我们对每一个半径 $r \in R$ 中的 $|H|$ 数值执行两种类型的标准化方法。第一种是和 hashtag 标签无关的标准化，方法是我们标准化每一个计数值，从而标准化半径 r 内所有 hashtag 标签的总计数值数。这个标准化给我们的一个想法，将每一特定的 hashtag 标签出现的相对频率与出现在该地区的其他的 hashtag 标签联系起来，从而标准化该地区的照片密度。二是规范化 hashtag 标签，在这里按我们从 Instagram 获得的带有特定的 hashtag 标签的图片总数目来规范化每一个计数值。该标准化方法使我们产生了一个新的想法，相对于特定概念的相对频率与这个概念出现的频率在整个美国的关系。我们进行了上述两种的标准化方法并且将特征向量联合起来构成了最后的特征向量，维数有 $2 \times |H| \times |R|$。

在我们的实验中，我们让 $C = H$，使用该组的类为一组 hashtag 标签来使问题变得简单一点，并且设定一组半径值 R ={1000, 2000，…，10000}。为了节省运算时间，我们量化所有的 GPS 坐标到 25000x50000 网格，结果就是每个方形网格单元覆盖 100×100 的区域。

3.6. 视觉语境特征

视觉语境特征类似于标签背景功能，除了在我们想利用围绕我们的 GPS 坐标 $(long_i, lat_i)$ 的视觉信号，而不仅仅是已经标记的 hashtag 标签的情况下。为了做到这一点，我们从各种在线的带有 GPS 坐标的社交网站上检索图像，并为每个图像运行一个和[26] 具有类似结构的 CNN，来为在互联网上出现的一些普通类型的 594 个概念计算概率，如“衣服”，“姑娘”和“咖啡”。594 个概念的列表在补充材料中有给出。

类似于标签背景特征，我们通过求和所有的图像落入所述半径的概率，从而得到 $(long_i, lat_i)$ 周围每个半径 $r \in R$ 的概率，对于每个独立的概念，得到一组数目为 $594 \times |R|$ 的概率值。然后，我们进行了之前提到两个类型的标准化方法，并连接形成了最终的特征矢量，是一个 $2 \times 594 \times |R|$ 维特征向量。我们使用了和标签背景特征相同的半径 R 和 GPS 的网格量化标准。

4. 学习最优池半径

在之前的一节里，我们介绍了标签背景特征和视觉语境特征。对于这些特征，我们解释了如何从汇总的hashtag标签和概念概率构建特征，通过串连在一起的归一化直方图来汇集在一组半径R。但是我们不指望所有的半径都是有益的，例如，对于hashtag标签和概念来说，半径就是罕见的，因为甚至是几公里远也可以成为主要指标。同样的，某些常用的hashtag 标签可能需要非常接近位置的精确定位。

半径学习层： 为了解决这个问题，我们将展示如何构建在 CNN 的一个层，它会自动学习用于池最佳的半径，我们称之为半径学习层。学习最佳半径是在许多方面是有潜在价值的。首先，通过集中在重要的最具价值的特征半径，可避免过度拟合。其次，我们可以想象，我们学习后的半径，可以让我们更清晰观察到CNN学习的是什么。首先，我们考虑一个hashtag标签/概念h的半径，适合一个函数$H_{(long_i, lat_i)}$，基于直方图的$h(\rho)$返回由$(long_i, lat_i)$确定的位置包括 hashtag 标签h和半径ρ对应的特征直方图的值。有几种方法，可以适应这样的功能，但为了简化计算，我们使用上一节R上计算的直方图值，并分段线性逼近得到的值。我们这样做是为每个训练的图像i所有的 hashtag 标签，概念，与这两种类型的归一化方案，获得一组$2\cdot(H+594)$直方图函数。

这些直方图函数的输出被视为代替级联直方图作为 CNN 的特征输入，每个 hashtag 标签/概念的半径参数ρ_h，被选择作为函数的输入值输入到神经网络。当计算反向传播的梯度时，我们将反向传播的梯度误差E计算进直方图函数H的梯度：

$$\frac{\partial E}{\partial \rho_h} = \frac{\partial E}{\partial H_{(long_i, lat_i),h}(\rho_h)} \frac{\partial H_{(long_i, lat_i),h}(\rho_h)}{\partial \rho_h} \quad (1)$$

在 RHS 的第一项是误差从半径学习层上面的网络架构导数传播到半径学习层，然后第二项是直方图函数$H_{(long_i, lat_i),h}$在ρ_h处的导数。由于我们使用分段线性逼近来配合直方图函数，第二中方法是比较容易计算的，通过计算在R上两个最近点之间的斜率。尽管我们可以实现更复杂的功能，但是我们发现线性逼近得到的梯度计算在所有训练例子的梯度计算中已经比较迅速和充分。由于考虑到 hashtag 标签/概念可能在有效半径之间的多个半径和加权信息，对于每个直方图函数，我们都重复了多次半径学习层。

5. 数据集

为了评估我们的方法，我们使用了最近发布的雅虎的Flickr Creative Commons 100M（YFCC 100M）数据集[4]，其中包含了版权来自Flickr版税的一

亿张图片。在这一亿张图片里，大概有 4,900 万张图片都带有 GPS 坐标的地理标注，这使得这个史无前例规模的数据集特别适合评估我们的方法。和图片被一同提供的还有 Flickr 自己编写的标签，我们使用它作为识别包含特定类的图像第一步骤。但是，因为该标签是非常嘈杂，我们必须手工验证和丢弃不属于我们感兴趣的类的图像。正如前面提到的，我们只专注于带有美国本土的地理标记图像。

选择位置敏感的类：其中一个我们要处理的问题是选择可能是位置敏感的类，并会从我们的位置上下文特征获益。这是非常重要的因为有的类是确定没有的，而且添加这些附加功能可能会导致分类器的过度拟合。实事求是地讲，我们需要一种方式来限制类的数量为一个我们可以注释并管理的数量。

为了解决这个问题，我们使用一个简单的数据驱动方法来选择类。使用大量 Instagram 的图像集合，我们估计了所有图像的离散地理空间分布 P，方法是通过将美国本土网格化，然后计算落入每个网格单元的图像数并标准化来创建有效的概率分布。然后，我们得到一个经常出现 Instagram hashtag 标签的一组类列表，对于每个类别 c，我们一类似的方式估计标记为类别 c 的图像的地理空间的分布 Q_c。有了这两个分布，我们比较了它们与库勒巴克-莱布勒（KL）散度的相似性：

$$D_{KL}(P||Q_c) = \sum_i P(i) \ln \frac{P(i)}{Q_c(i)} \qquad (2)$$

直观地说，我们想找到那些表现的和所有图像的分布不一样的地理空间分布的类，因为这将表明它们具有一些位置敏感性质。KL 散度通过给人的两者之间的概率分布差异的量度做到了这一点，我们选择最高 KL 散度的前 100 个类。在实践中，当被给了一个新的类别 c，我们可以简单计算 $D_{KL}(P\|Q_c)$ 和阈值以确定该类是否将从我们附加的位置特征中受益。地理空间分布的例子如图 5 所示。

YFCC100M-GEO100 数据集：利用被选择的具有最高 KL 散度值的前 100 个类别，我们手动验证并注明了 YFCC100M 中一个很大的被 Flickr 的用户使用类别大肆标记的图像集。结果是一个 88,986 的图像的数据集，意味着至少每类 100 张图片，我们把它记为 YFCC100MGEO100 数据集，并会公开。我们所选择的类别有物体，有地点还有场景，例如“秋天”，“海滩”还有“鲸鱼”，说明了我们努力去做的分类多样性。图 6 为在数据集中的图像的可视 GPS 坐标分布。类的完整列表在附加材料中给出。

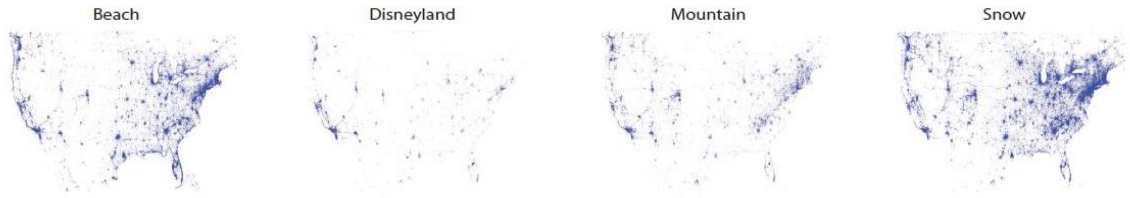


图5:在美国本土不同类别的Instagram的hashtag标签分布。虽然我们可以看到有趣的图案,例如尽量靠近海岸沙滩hashtag标签和阿巴拉契亚山脉的山 hashtag 标签的轮廓，但还是有很大噪音。

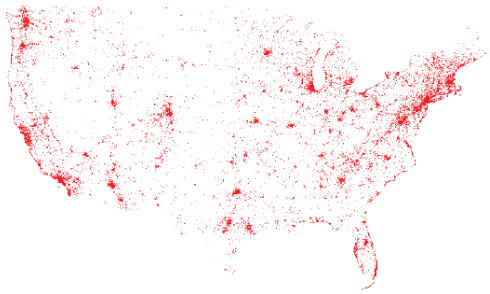


图6:我们介绍的 YFCC100M-GEO100 数据集中所包含的88,986张图片的地理分布

Method	Mean AP	Acc@1	Acc@5
Image only	36.82%	39.45%	70.15%
Image + GPS coordinates	36.83%	39.47%	70.23%
Image + GPS encoding 10x20	38.58%	41.48%	72.39%
Image + GPS encoding 100x200	**38.89%**	**41.67%**	**72.47%**
Image + Geographic map feature	37.70%	40.28%	70.79&
Image + ACS feature	**40.41%**	**42.79%**	**73.84%**
Image + Hashtag context feature	39.86%	42.27%	73.38%
Image + Visual context feature	38.81%	41.53%	72.31%
Image only (SVM)	33.41%	36.56%	60.05%
Image + All features (SVM)	34.61%	38.06%	62.88%
Image + All features χ^2 kernel (SVM)	**35.12%**	**38.57%**	**63.74%**
Image + Flickr prior 10NN	24.15%	25.36%	36.46%
Image + Flickr prior 100NN	33.38%	35.45%	60.62%
Image + Flickr prior 1000NN	**36.30%**	**37.86%**	**68.57%**
Image + Instagram prior 1000km	24.03%	22.70%	38.23%
Image + Instagram prior 4000km	31.96%	30.62%	58.69%
Image + Instagram prior 8000km	**33.08%**	**30.67%**	**60.13%**

表 1：比较不同基线方法的结果。对于 CNN 模型，我们不使用 pre-cat 和 post-cat 层。

6. 结果

我们将 YFCC100M-GEO100 数据集随机划分成 80%的训练集和 20%的测试集。我们进一步分出一小部分的训练集作为我们的模型参数调整时的验证集。

实现细节：根据[26]，我们使用 0.9 的动力和 0.005 衰变随机梯度下降培训我们的模型。我们使用 0.1 的学习率，并运行数据约 30 次，每 10 次降低 0.1 学习率。我们为我们所有的完全连接层使用了一个 0.5 丢弃比率。因为我们的训练数据是比较小的，我们通过使用一个含有大量 Instagram 图像的组进行预训练，并初始化在该模型的图像的网络部分中的参数（参见图 2），然后将预训练参数传进我们的模型。请注意，我们也可以进一步微调这些参数，但不会因为速度而更改参数。

性能指标：为了评估我们的模型，我们选用了三种不同的性能指标。除了平均精度（AP）的标准度量值，我们还包括了用于归一化的精确度@1 和归一化的精确度@5 的结果，因为它们在最近的论文[24]以及 ImageNet 分类挑战[35]中使用的动机。归一化的精确度@k 度量指示前 k 个预测的包含地面实况标签测试样品的级分，将每个类

进行归一化处理来适应每一类的图片数目的不同。

6.1. 基线方法

我们评估每个提出的特征的好处，表 1 中顶部的两个部分表示的是没有 pre-cat 层和 post-cat 层作为基线（参见图 2）。毫无惊奇的，使用 GPS 坐标不会产生任何性能显著增益，因为它们在线性分类器的上下文中毫无意义。使用GPS的编码特性，通过2%左右的所有性能测量的增益。我们可以得到更好的性能。我们也可以看到，对于每一个功能，我们用基线图像特征串联特征后获得的性能增益，这说明他们提供了补充信息。特别是，ACS 特征产生性能增幅最大，其中所有的性能度量值都有近 4%的涨幅。

支持向量机模型：我们使用支持向量机（SVM）和核回归（kernelizing）进行我们的实验。我们使用内核平均结合特征，正如它被认为和其他更复杂的方法同为执行标准[17]。在表 1 的中间部分，我们展示了使用多级铰链损失 SVM 分类和交叉验证调整参数的结果。对于这种简单的结合，我们对所有特征使用线性内核，对于 χ^2 组合，我们计算了直方图特征的 χ^2 内核（包括 hashtag 标签背景下，视觉环境），并出于维度的考虑，为剩余的也使用线性内核。一般情况下，我们发现 SVM 比 softmax 分类器的表现更糟。用 χ^2 内核进行直方图特征的核回归的方法表现出来的性能比仅使用线性内核更好，但仍没有超出 softmax 的性能。

贝叶斯先验：根据[9]中所用的方法，我们还试图结合位置上下文作一个贝叶斯先验。使用贝叶斯法则，预测类别 c 的给定图像 I_i 和位置 $(long_i, lat_i)$ 概率可以写成：

$$P(c|I_i, long_i, lat_i) = \frac{P(I_i, long_i, lat_i|c)P(c)}{P(I_i, long_i, lat_i)} \quad (3)$$

假设图像和位置条件独立于给出的类别，进一步运用贝叶斯规则并删除不依赖于 c 的组术语，我们得到：

$$P(c|I_i, long_i, lat_i) \propto \frac{P(c|I_i)}{P(c)} P(c|long_i, lat_i) \quad (4)$$

在我们的实验中，我们假设 P_c 一个在班前均匀。我们尝试了两种不同的方法计算位置先验 $P(c|long_i, lat_i)$，结果在表 1 的底部两个部分给出。在 Flickr 先验，对于每个测试图像，我们可以根据 GPS 可见训练组找到 k 个最近邻位置（k 近邻），并使用他们的标签来估计位置的先验。在 Instagram 先验，我们对每个测试图像，都使用了一定半径 r 内由 hashtag 标签特征计算出的直方图，并使用标准化后的直方图作为位置先验的分布。虽然任何一种方法都没有改进总体结果，但是有趣的是，对于某些类，如“迪斯尼乐园”，结果是这两种类型的先验

概率的平均 AP 有超过 45%的改善。然而，对于大多数类别，位置先验受损多于受益，造成性能的整体下降。

Method	Mean AP	Acc@1	Acc@5
Image only	36.82%	39.45%	70.15%
Image + All features with -/-	37.97%	40.19%	70.67%
Image + All features with 128/-	42.22%	44.76%	75.74%
Image + All features with 256/-	42.34%	**44.82%**	**75.86%**
Image + All features with 512/-	42.20%	44.43%	75.53%
Image + All features with 1024/-	41.60%	43.98%	75.16%
Image + All features with 256/4096	**43.28%**	43.74%	74.30%

表 2：所有的功能和不同的 pre-cat 和 post-cat 层串联后的结果。符号 X/ Y 中的 X,Y 分别表示的是 precat 层的维数和 post-cat 层的维数，而-表示的是没有 pre-cat 或 post-cat 层。

6.2. 特征组合结构

我们评估了各种结构的特征组合，评估了在模型中的级联层前后不同程度的深度水平产生的影响。结果在表 2 中有给出。表 2 的上半部分给出的是在每个特征的 pre-cat 层添加额外的深度的结果，底部展示的是使用一个 4096 维 post-cat 层的结果。我们将所有的都和表 1 中的“只有图像”的模式进行了比较，即我们现在所说的基准图像模型。

Pre-cat 层：从结果中我们看到，简单的将特征串联起来不会导致性能的显著增加，很可能是因为特征的维数太大，模型过度拟合。因此，我们引入 pre-cat 层来捕捉特征之间的关系，并用它来降维。虽然它们的表现类似，但是 256 维的层似乎在性能和参数的学习中获得了最佳平衡，在所有指标上都获得了几乎 6%的性能涨幅。我们也尝试在层之外增加额外的深度，但我们发现这样做并没有意义而且急剧增加了需要学习的参数的个数。

Post-cat 层：我们也进行了用 post-cat 层以捕捉不同特征之间关系的实验。我们发现，一个 4096 维的完全连接层似乎会略微提升平均 AP，但是会由于过度拟合导致两个标准化准确率的下降。再次如先前观察到的，在这里加入额外的深度也使模型过度拟合，降低了模型的性能。

Method	Mean AP	Acc@1	Acc@5
Image only	36.82%	39.45%	70.15%
Image + Hashtag context feature	39.86%	42.27%	73.38%
Image + Hashtag context feature RL5	40.19%	42.52%	73.57%
Image + Hashtag context feature RL10	**40.80%**	**43.10%**	**74.15%**
Image + Visual context feature	38.81%	41.53%	72.31%
Image + Visual context feature RL5	38.75%	41.31%	72.08%
Image + Visual context feature RL10	**39.07%**	**41.78%**	**72.48%**
Image + All features with 256/-	42.34%	44.82%	75.86%
Image + All features with 256/- RL10	42.91%	**45.17%**	**76.09%**
Image + All features with 256/4096	43.28%	43.74%	74.30%
Image + All features with 256/4096 RL10	**43.78%**	44.14%	74.70%

表 3：学习最佳池半径后的结果。RL5 和 RL10 指用于替换级联直方图所用的半径学习层的数量（5,10）。

6.3. 学习最佳池半径

在前面的章节中，我们串联了在标签背景环境和视觉语境特征的各种半径条件下计算的直方图。由于在有效半径中经常有多个半径和权重，我们用多个半径学习层替换串联直方图，结果如表 3 所示。在顶部的两部分，我们发现标签背景环境有大幅提升，以及没有 pre-cat 和 post-cat 层的视觉语境特征略有提升。在最底部的部分，通过为上一节我们的最佳模型使用半径学习层，我们获得了 0.5%的平均 A P 额外涨幅。我们再次发现，添加的 post-cat 层导致模型轻微的过度拟合，因此在下面的分析中，我们

使用 256/- RL10 模型作为最佳模式。 图 7 展示了一些有趣的例子和预测。

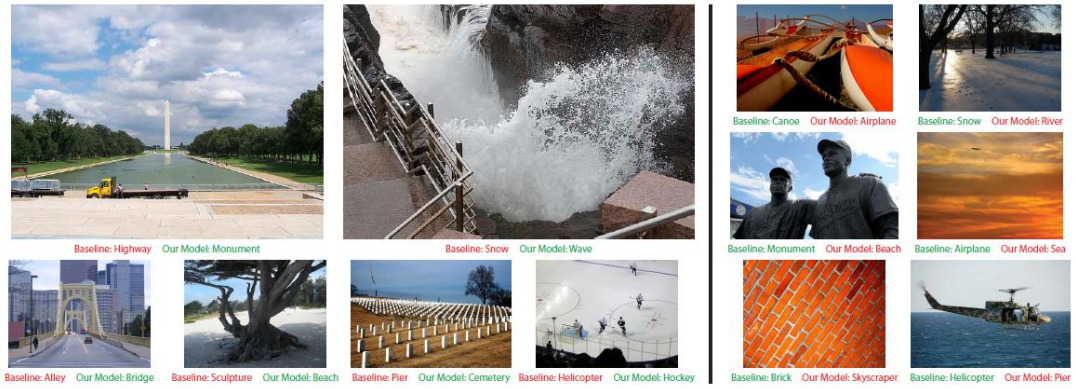


图 7：基线图像模型和我们的最佳模式（256/ - RL10）比较的示例结果，正确的预测表示为绿色，不正确的预测则为红色。在补充材料中有给定的图像分值

6.4. 学习最佳池半径

在前面的章节中，我们串联了在标签背景环境和视觉语境特征的各种半径条件下计算的直方图。由于在有效半径中经常有多个半径和权重，我们用多个半径学习层替换串联直方图，结果如表 3 所示。在顶部的两节中，我们看到了标签背景特征的大幅提升，以及没有 pre-cat 层和 post-cat 层的视觉语境功能轻度提升。在最底部的章节，通过为上一节我们的最佳模型使用半径学习层，我们得以获得额外的 0.5%的平均 AP 涨幅。我们再次发现，添加的 post-cat 层导致模型轻微的过度拟合，因此在下面的分析中，我们使用 256 / - RL10 模型作为我们的最佳模式。图 7 展示了一些有趣的例子和预测。

最佳和最差类别：在图 8 中，我们展示了和基准图像模型相比较，20 个表现最好和 20 个表现最差的类别。具体的位置类别，像“迪斯尼乐园”，“赌场”，“恶魔岛”这种在性能上有较大的增加，因为它们是局限于一个或少数的美国地点。在另一方面，一些表现较差的类别是汽车品牌，这表明细粒度车类不是非常具有位置特定性，或是不太适合我们的特征和模型。但是，由于我们选择位置敏感概念的方法是数据驱动和非监督，所以他们都包括在内。

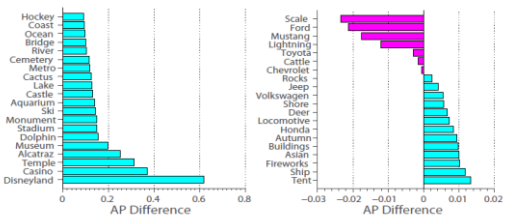


图 8：我们最好的模型（ 256/- RL10）中 20 最佳和最差的类和基准图像模型之间的 AP 差异。

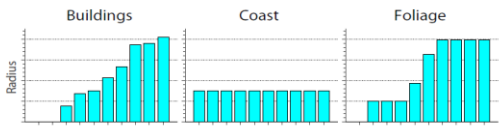


图 9：我们的最佳模式（256/ - RL10）中，按照从最小到最大排列的三类类别的可视化学习后的半径。

学习后的半径参数：我们从图 9 中可以看到一些类别学习后的半径参数。我们发现对于大多数概念，半径学习层的 10 个不同的副本通常收敛到 3 个或更少的不同的半径，像“海岸”和“叶子”这种类，这表明某些半径确实提供更多的信息。有时，某些类，如“建设”学习几乎所有不同的半径，可能是因为城市地区建筑物的丰度使小半径重要，而农村地区大半径更重要。

7. 结论

在本文中，我们引入了利用位置环境进行图像分类的问题。为了表示位置环境，我们提出了 5 个特征，可以帮助捕捉一个特定位置的背景，并展示如何将它们纳入一个 CNN 模型。对于需要集中半径的特征，我们将展示如何在同一框架内自动学习最佳半径，使我们能够获得更好的性能和更深入的了解进入网络参数。此外，我们引入并公开提供了 YFCC100M-GEO100 数据集，这是我们通过手动标注以获得类标签地理标记的图像。

对于今后的工作，我们想探索利用现在越来越常见的图像其它方面的特征，如时间和拍摄日期或用户之间的社会关系。

参阅资料

[1] American community survey. http://www.census.gov/acs/www/. 4

[2] Google maps. http://maps.google.com. 3

[3] Instagram. http://www.instagram.com. 4

[4] Yahoo flickr creative commons 100m. http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67. 2, 5

[5] K. Amlacher, G. Fritz, P. M. Luley, A. Almer, and L. Paletta. Geo-contextual priors for attentive urban object recognition. In ICRA, 2009. 2

[6] S. Ardeshir, A. R. Zamir, A.Torroella, and M. Shah. Gisassisted object detection and geospatial localization. In ECCV, 2014. 2

[7] G. Baatz, O. Saurer, K. K¨oser, and M. Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *ECCV*, 2012. 2

[8] M. Bansal and K. Daniilidis. Geometric urban geo-localization.

In *CVPR*, 2014. 2

[9] T. Berg, J. Liu, S. W. Lee, M.L. Alexander, D. W. Jacobs,and P. N. Belhumeur.
Birdsnap Large-scale fine-grained visual categorization of birds.
In *CVPR*, 2014. 2, 7

[10] A. Bergamo, S. N. Sinha, and L. Torresani.
Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification.In *CVPR*, 2013. 2

[11] L. Cao, J. Yu, J. Luo, and T. S. Huang.
Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In *MM*, 2009. 2

[12] S. Cao and N. Snavely.
Graph-based discriminative learn-ing for location recognition.
In *CVPR*, 2013. 2

[13] D. M. Chen, G. Baatz, K. Koser, S. S. Tsai, R. Vedantham, T. Pylv an ainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices.
In *CVPR*, 2011. 2

[14] D. J. Crandall, L. Backstrom, D. P. Huttenlocher, and J. M. Kleinberg.
Mapping the world’s photos.
In *WWW*, 2009. 2

[15] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert.
An empirical study of context in object detection.
In *CVPR*, 2009. 2

[16] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros.
What makes paris look like paris?
SIGGRAPH, 2012. 2

[17] P. V. Gehler and S. Nowozin.
On feature combination for multiclass object classification.
In *ICCV*, 2009. 7

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik.
Rich feature hierarchies for accurate object detection and semantic segmentation.
In *CVPR*, 2014. 2

[19] P. Gronat, G. Obozinski, J. Sivic, and T.
Pajdla. Learning and calibrating per-location classifiers for visual place recognition.
In *CVPR*, 2013. 2

[20] J. Hays and A. A. Efros.
IM2GPS: estimating geographic information from a single image.
In *CVPR*, 2008. 2

[21] A. Irschara, C. Zach, J. Frahm, and H. Bischof.
From structure-from-motion point clouds to fast location recogni-ti

on.
In *CVPR*, 2009. 2

[22] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless.
Geolocating static cameras.
In *ICCV*, 2007. 2

[23] D. Joshi and J. Luo.
Inferring generic activities and events from image content and bags of geo-tags. In *CIVR*, 2008. 2

[24] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei.
Large-scale video classification with convolutional neural networks.
In *CVPR*, 2014. 6

[25] A. Khosla, B. An, J. J. Lim, and A. Torralba.
Looking beyond the visible scene.
In *CVPR*, 2014. 2

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton.
Imagenet classification with deep convolutional neural networks.
In *NIPS*, 2012. 1, 2, 3, 4, 6

[27] S. Lee, H. Zhang, and D. J. Crandall.
Predicting geo-informative attributes in large-scale image collections using convolutional neural networks.
In *WACV*, 2015. 2

[28] 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. 2

[29] D. Leung and S. Newsam.
Proximate sensing: Inferring what-is-where from georeferenced photo collections.
In *CVPR*, 2010. 2

[30] A. Li, V. I. Morariu, and L. S. Davis.
Planar structure matching under projective uncertainty for geo-location.
In *ECCV*, 2014. 2

[31] Y. Li, N. Snavely, and D. Huttenlocher.
Location recognition using prioritized feature matching.
In *ECCV*, 2010. 2

[32] T. Lin, S. Belongie, and J. Hays.
Cross-view image geo-localization.
In *CVPR*, 2013. 2

[33] K. Matzen and N. Snavely.
Nyc3dcars:
A dataset of 3d vehicles in geographic context.
In *ICCV*, 2013. 2

[34] R. Raguram, C. Wu, J. Frahm, and S. Lazebnik.
Modeling and recognition of landmark image collections using iconic scene graphs.

IJCV, 95(3):213－239, 2011. 2

[35] O. Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge, 2014. 2, 6

[36] T. Sattler, B. Leibe, and L. Kobbelt. Improving image-basedlocalization by active correspondence search. In *ECCV*,2012. 2

[37] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *CVPR*, 2007. 2

[38] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *IJCV*, 2008. 2

[39] C. Szegedy et al. Going deeper with convolutions. *CoRR*, 2014. 2

[40] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In *CVPR*, 2013. 2

[41] A. Torralba. Contextual priming for object detection. *IJCV*,53(2):169－191, 2003. 2

[42] J. Yu and J. Luo. Leveraging probabilistic season and location context models for scene understanding. In *CIVR*, 2008.2

[43] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *ECCV*, 2010. 2

[44] Y. Zheng et al. Tour the world: Building a web-scale landmark recognition engine. In *CVPR*, 2009. 2

[45] B. Zhou, L. Liu, A. Oliva, and A. Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *ECCV*, 2014. 2