Lightweight Unsupervised Deep Loop Closure

Nate Merrill and Guoquan Huang

Robotics: Science and Systems

韩立明

Abstract

- 鲁棒有效的闭环检测对于大规模实时SLAM至关重要。在本文中,我们提出了一种新的**无监督深度神经网络体系结构**,该体系结构用于视觉闭合的特征嵌入,既可靠又紧凑。我们的模型基于自动编码器架构,专门针对当前的问题量身定制。为了训练我们的网络,我们像去噪自动编码器一样对我们的输入数据**施加随机噪声**,但是,我们不是应用随机丢失,而是使用随机**投影变换来扭曲图像**以模拟由于机器人运动引起的自然视点变化。(数据增强)
- 此外,我们利用由定向梯度直方图(HOG)提供的几何信息和光照不变性,使编码器重建HOG描述符而不是原始图像。因此,我们训练好的模型直接从原始图像中提取出对外观极端变化的强大功能,而无需标记的训练数据或环境特定的训练。我们对各种具有挑战性的数据集进行了大量实验,表明所提出的深度环闭合模型在有效性和效率方面始终优于最先进的方法。我们的模型快速可靠,足以实时关闭环路而不降低维数,并且能够在最先进的基于ConvNet的环路闭合系统中取代通用的现成网络。

INTRODUCTION

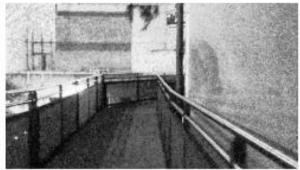
- 在线闭环检测作用: vslam减小累计误差
- 观察角度、照明、天气、动态目标,闭环 检测具有巨大挑战
- ConvNets在视觉分类上取得巨大成就,但在 place recognition上特征提取慢,查询慢, 需要大量带标签的数据训练

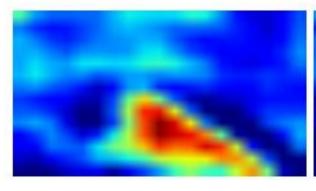
INTRODUCTION

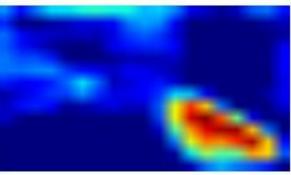
来自Gardens Point数据集的示例图像匹配,其展示了观察角度,动态对象和照明以及遮挡的巨大差异。

然而,使用正确的图像作为查询,我们提出的方法在我们的实验中正确地检索左边(参见第IV节),而所有测试的最先进的方法都检索不正确的图像。 在每个图像下方,给出了在展开之前描述符层的第一级。 显然,这些视觉上不相似的图像被转换成非常相似的激活图。









INTRODUCTION

了解决上述问题, 在本文中, 我们构建了一个新 的基于自动编码器的ConvNet for loop closure, 需要**很少的参数**,并以**无监督**的方式使用公共数据进行训练。特别是,在构建我们的自动编码器网络 我们利用经典几何视觉技术的优势 - 定向梯度 直方图(HOG)[19]提供了一种在保留显着特征的 同时压缩图像的便捷方法,以及投影变换(单应性) [20]将具有不同观点的图像联系起来。相比之下, 我们还将现代堆叠卷积自动编码器整合到网络中以 因此, 从我们的网络中提取的特征 不仅对外观的极端变化具有鲁棒性,而且对于实时 循环闭合而言也足够轻巧和紧凑 - 即使对于资源受 限的系统也是如此。

contributions

- 本文的主要贡献如下:
- 我们设计了一种无监督的卷积自动编码器网络架构,专为闭环检测而设计,并且适用于高效,稳健的场所识别。
- 我们对提出的深环闭合模型与不同数据集上的最新方法进行了广泛的比较研究。为了使研究界受益,我们开源我们的代码和预先训练的模型以及一个新的数据集,该数据集捕获单个数据集中的视点,天气,光照和动态对象的极端变化。
- 本文的其余部分结构如下:在第Ⅲ节中,我们详细介绍了所提出的深度闭环网络,包括网络架构,训练方案和在线使用。所提出的方法在第Ⅳ节中进行了广泛测试-既针对最先进的算法,也针对实时循环闭包设置。最后,我们在Section中总结了这篇论文

RELATED WORK

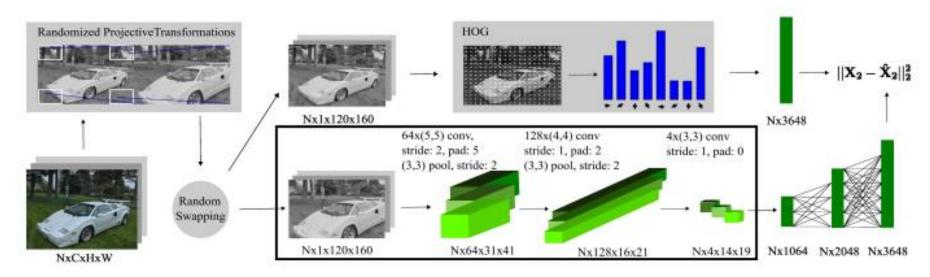
• Bow FAB-MAP: 大的变化 fail

• ConvNet: 复杂描述子慢

• NetVLAD: 标签数据

- Design Motivation
- 使用单应矩阵变换的合成视点改变测试重定位系统,我们采用更广义的视点改变(投影变换)来训练我们的深环闭合模型。
- HOG描述符是相同大小的图像的固定长度矢量, 并且可以自然地通过**欧几里德距离**进行比较,
- 由于HOG依赖于梯度方向,因此对于视点的改变不是非常鲁棒,但另一方面,图像梯度在某种程度上是光照的反映。

- Network Architecture
- N是仿射变换产生的图像对数量
- 仿射变换替代labeled
- 使用4*14*19作为特征输出



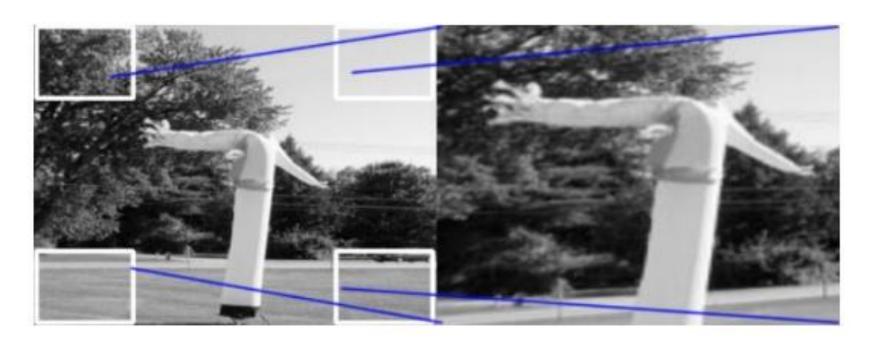
- Network Architecture
- 上图提供了从原始图像到损耗层的数据流。在训练开始之前,训练图像集I中的每个图像被转换为灰度,调整为120×160,并用于**创建图像对**。针对来自每对的随机选择的图像计算HOG描述符。我们堆叠来自每批训练图像的所有HOG描述符,由维度N×D的X2表示,其中N是批量大小,D是每个HOG描述符的维度。该对中的另一个图像保持原始形式,并与该训练批次中的其他N-1个图像一起堆叠。由X1表示的所得张量具有N×120×160的尺寸。
- 训练网络旨在仅

使用两个卷积和pool层,一个纯卷积层和三个完全连接层来重建给定X1的X2。请注意,每个图层都有一个激活后。我们对卷积层使用(ReLU)激活,而为完全连接的层选择S形激活,以便更好地重建HOG描述符(因为它将数据标准化为[0,1])。另外,由于欧几里德距离自然是HOG描述符的良好距离度量,我们使用2范数损失函数来比较X2与其重建X2。在应用时,我们只保留X1和三个卷积层。与最先进的位置识别模型[16,15,12]相比,我们的模型非常轻巧,仅占用139 MB的GPU内存,为其他流程提供了充足的空间。

- Network Training
- 如前所述,所提出的模型不需要标记训练图像或包含任何特定信息-也就是说,可以在训练集中使用来自任何场景的任何图像来改进我们 的模型。 为了说明这一点,我们在Places数据集[37]上训练了我们的 模型, 该数据集有超过800万张图像, 最初是为场景识别而设计的。 图2和3包含来自该数据集的一些图像示例。 虽然Places数据集中的大 多数图像与闭环检测算法可能遇到的任何场景无关,但是**相对于较小** 数据集的训练,图像的绝对数量可以提高性能。算法1概述了利用这 样的数据集来创建T1和T2的主要步骤,T1和T2是在每次随机梯度下 降迭代期间从中采样X1和X2的张量。

Network Training

• 仿射变换 给定一个图像I∈I,我们想自动生成Iw,它与我在不同视点中的场景相同;通过将**随机二维投影变换矩阵Hp**∈R3×3应用于I中的每个像素位置来实现该效果。为了获得该矩阵,沿着图像I中的角在边界框内随机选择四个点(见图3);然后计算Hp以使I变形,使得这四个点成为Iw的四个角。 我们选择点选择的每个边界框为H/4×W/4,以避免Iw的过度失真,同时仍然使其变形足以模拟场景的新视角。产生H矩阵(不考虑深度)



- Online Use
- 一旦我们的模型被训练,在部署到在线使用时,我们创建一个由我们的模型 提取的描述符的数据库,然后查询它以找到循环闭包候选者。虽然K-D树[40] 是为最近邻搜索创建此类数据库的常用方法,但对于1,064维向量的线性搜索 没有加速 - 即使搜索是近似的[41]。因此,我们使用简单的线性搜索方法。另 外,由于描述符足够紧凑,因此可以直接计算它们的相似性而不降低维数。
- 我们试图强调,我们使用从模型中提取的描述符创建和查询数据库的方法简单但有效;尽管如此,我们能够以最少的内存使用率实现快于实时的查询速度(参见第IV-E节)。此外,由于许多新的基于ConvNet的场所识别系统 [12,32,14,33]依赖于笨重的现成网络的特性,我们的轻量级模型可以在许多这些系统中使用,以实现加速竞争准确性(见第IV-G节)。

- Example
- 为了验证所提出的无监督深度环闭合模型,我们对各种数据集进行了广泛的 比较研究,其中包括最先进的方法以及适用的其他基准。运行时被用作评估 效率的标准,我们利用**精度和召回曲线,一种评估二元分类的标准方法**,来 量化有效性。虽然有许多方法可以解释精度和召回曲线,但我们主要使用:
 - (i) 曲线下面积(AUC),其中需要更高的AUC;(ii)具有100%精度的最大召回率,用r表示,其中再次需要更高的值。这可以在任何精确回忆曲线中直观地观察到,因为它将是精度首次从1.0下降的召回率。通过观察这两个值,我们可以全面了解所考虑的算法可以概括的程度;然而,r值在实践中稍微更理想,因为尽管AUC高,但一个二元分类器对于所有召回率可能具有非完美的精度。

- 精度:正确回环检测占总检测回环数目的比率(表明误报)
- 召回:正确回环检测占测试样本总回环的比率(表明漏报)
- SLAM对准确率更严格,对召回率则宽松一点。因为假阳性(检测是而实际不是)回环会在pose graph中添加错误的边,优化算法可能会给出完全错误的结果。想象一下,slam程序把所有的办公桌当成同一张,则走廊不直了,墙壁被交错在一起了,最后整个地图失效。召回率低一些,则顶多有部分的回环没有检测到,地图可能会受一些累积误差的影响,然而仅需一两次回环就可以完全消除它们了。

算法/事实	是回环	不是回环
是回环	真阳性(True Positive)	假阳性(False Positive)
不是回环	假阴性(False Negative):	真阴性 (True Negative)

$$Precision = \frac{TP}{TP + FP}$$

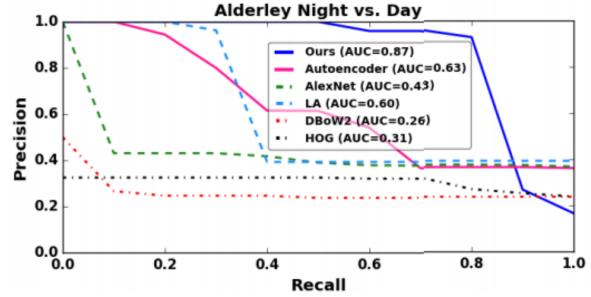
$$Recall = \frac{TP}{TP + FN}$$

- Example
- 我们将提出的方法与以下内容进行比较: 它不是重建生成图像的HOG描述符, 和解开层来重建原始图像并且在训练期间受到随机丢失。 其效率与我们的无监督模型相当, [16], 需要标记数据进行训练(即监督),使任何再训练变得困难。 我们使用来自最先进的ORB-SLAM的DBoW2词汇树[24,23]。 Sünderhauf等。 [11]发现AlexNet conv3是最强大的地点识别层;然而, 产生的64,896维向量太大而不能执行实时数据库查询。 的高斯随机投影(GRP)[42,43]来将conv3层压缩到与所提出的模型中的描述 在我们的测试中,我们使用BVLC培训的AlexNet。 虽然3,648维HOG描述符用于训练我们的模型, 但我们将其包括在内比较仅 是为了表明我们的模型能够比原始的简化HOG学习更好的特征, HOG作为地点识别描述符的能力。 对于所有这些 以便纯粹比较每个描述符匹配位置的能力。 出的是,在以下所有实验中,我们的方法使用在与测试数据集完全不同的 仅据集上训练的相同模型,表明所提出的深环闭合网络不需要特定于环境的

The Alderley Dataset



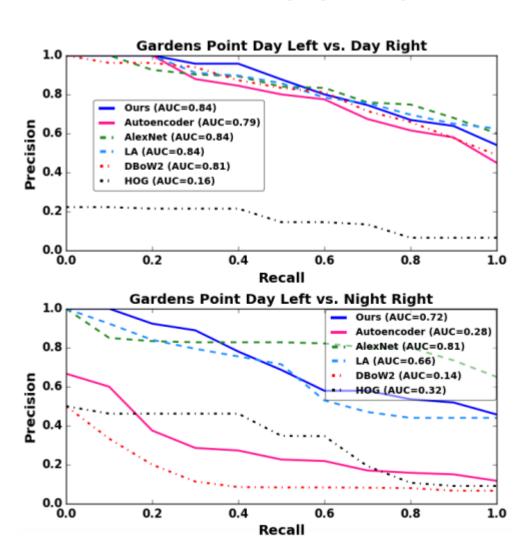




- The Alderley Dataset
- Alderley数据集最初是在SeqSLAM [2]中引入的,由两个图像序列组成,从雨夜和晴天拍摄的视频中提取。图4显示了来自该数据集的示例图像匹配;即使是人类也很难意识到这些图像是同一个地方。
- 帧对应包含在数据集中,为地点识别提供基础事实,并为相同位置的多个连续帧增加容差。我们测试每个序列的最后200帧。比较结果显示在图5中。显然,在这种情况下,我们的方法是最强大的。有趣的是,常规自动编码器在这里表现良好。请注意,[16]中的模型是在Alderley数据集的不同子集上训练而不是在此处使用,使得他们的模型优于未在任何Alderley数据集上训练的其他模型。尽管如此,我们的模型在这个实验中仍然更加稳健,而其他方法由于该数据集中外观的显着差异而失败。

- The Gardens Point Dataset
- Gardens Point数据集包括通过澳大利亚布里斯班QUT校园的三次遍历。在此数据集中,有两个白天路径 一个往往包含行走路径左侧的图像,另一个包含右侧。此外,还有一个夜间遍历,也趋向于路径的右侧。与Alderley数据集不同,来自一个序列的图像来自其他两个中的任何一个。我们利用这一点,以及对同一位置的多个连续图像的额外容差,除了本工作中剩余的精确回忆实验之外,还定义了该实验的基本事实,因为其余的数据集遵循相同的格式。
- 该数据集的一个例子如图1所示,而图6显示了比较结果。图6(上)是白天左视和白天右视序列。我们的方法,AlexNet和LA在这个数据集中表现相当,甚至自动编码器也不甘落后;但是,我们将看到这种具有可比性能的趋势在整个实验中都没有。DBoW2在这个实验中具有竞争力,
- 但是没有达到我们的方法,AlexNet和LA。我们用来训练模型的HOG描述符显然不如我们的最终描述符那么健壮,即使在这个目间数据集中 实验中使用的更容易的数据集之一。为了进一步挑战这些方法,我们使用Gardens Point数据集中的夜间序列 其结果如图6(下图)所示。在这种情况下,我们的方法采用highestr值和第二高的AUC。DBoW2,HOG和自动编码器在此测试中完全失败。尽管通过在夜间图像上训练ORB词汇树,很可能可以提高DBoW2的性能,但我们希望测试所有这些方法,而不是为了概括而进行任何特定于环境的训练。

The Gardens
 Point Dataset



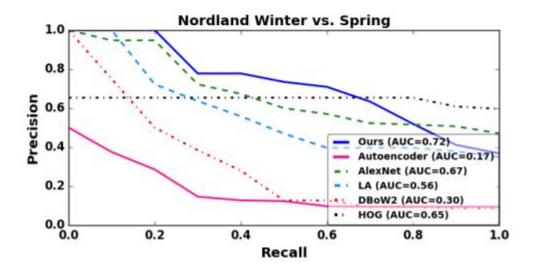
- The Nordland Dataset
- Nordland数据集是迄今为止最具挑战性的地方识别数据集之一 包括四个时间同步的挪威火车旅行视频。四个9小时长序列中的每 对应于不同的季节,对跨季节地点识别产生了 集的图像对的示例。 Winter vs. Spring。具体来说, 实验使用框架29至200进行, 生运动并且在隧道之外。 导致精确回忆曲线中 该注意Lopez-Antequera等。 |的最后一小时外,其余都使用了 己经在训练阶段看到了这些测试数据。 我们的模型也优于他们的模型,

The Nordland Dataset





Fig. 7: An example image pair from the Nordland dataset. The left image is from the spring sequence while the right one is from the winter.



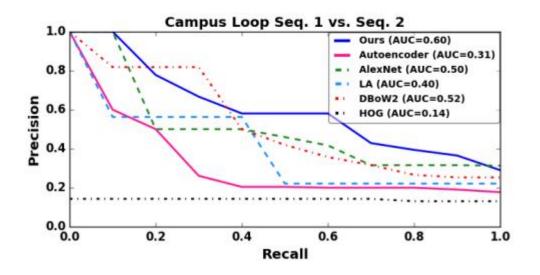
- Our Campus Loop Dataset
- Nordland数据集提供极端天气变化,Gardens Point数据集提供极端亮 而Alderley数据集提供 发现没有数据集可以提供所 来自Lopez-Antequera等人的模型。 表现明显差于其他三种深度学习 中甚至达不到DBoW2。

Our CampusLoop Dataset





Fig. 9: An image pair example from our Campus Loop dataset, which has extreme variations in viewpoint, weather, and dynamic objects.

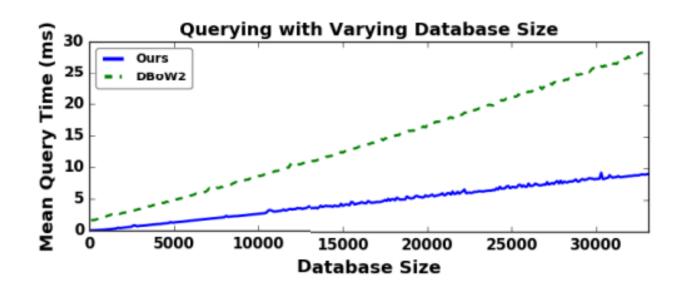


- Runtime Evaluation
- 为了验证我们的方法的效率,我们对描述符计算时间和单个最近邻搜索的数据库查询时间执行运行时评估。这些测试在价格合理的硬件上进行,以实现更好的可重复性 特别是i7-6700HQ CPU和GeForce GTX 960M GPU。请注意,在此测试中,我们仅针对DBoW2和AlexNet进行比较,原因如下:
- (i) Lopez-Antequera等人。[16]不提供用于执行图像匹配的开源库,因此在我们的代码中对其模型计时并不一定公平。他们使用GPU报告1.8毫秒的描述符计算时间,这比我们的慢。然而,它们的描述符比我们的描述符小,因此查询应该更便宜,而我们的准确性显示出竞争力(如果不是更好),并且更方便微调或重新训练。(ii)在前面的测试中显示减少的HOG只是为了表明我们的模型学习了更好的版本,使其运行时无关紧要。我们选择具有ORB特征的DBoW2作为一个基准,因为它是许多最先进的SLAM系统中使用的最快的位置识别库之一(例如,[23,24])。另外,我们选择具有GRP的AlexNet conv3功能,压缩到1,064维度,以及原始形式,因为它是基于ConvNet的场所识别的流行选择(即[12,14,33,32])。请注意,AlexNet已被修改为仅包含conv3层,以进行公平测试。

- Runtime Evaluation
- KITTI Visual Odometry数据集序列00 [44]用作第一个实验的测试数据。我们利用这个序列中的4,541个376×1241立体对来构建两个子集,将所有左图像放在数据库中,并使用正确的图像进行查询。表I显示了该实验的结果,其中特征提取时间是指从原始图像开始到将该图像的表示插入到相应数据库中结束的时间。
- 查询时间不包括任何描述符计算时间。请注意,DBoW2没有GPU实现,带有GRP的AlexNet会生成与我们相同大小的功能,因此查询时间将相同。从表I可以清楚地看出,当GPU用于通过网络进行前向传递时,我们的方法比其他方法更快,并且在使用CPU时仍然相当快。此外,我们的查询方法虽然很简单,但在本实验的速度方面优于DBoW2。我们还测试了可变大小数据库的查询速度,仅比较了DBoW2--表I中最具竞争力的候选者。我们使用大型圣卢西亚数据集[45],类似于KITTI,是一系列立体声对。但是,此序列包含超过30,000个立体对,这使其非常适用于测试可变数据库大小。左图像用于数据库,右图像的子集用于查询。图11显示了该实验的结果,从中可以看出,即使对于非常大的数据库,我们的查询方法也很便宜-大于典型SLAM系统创建的数据库。

Runtime Evaluation

Method	Extract (GPU)		Extract (CPU)		Query	
	μ	σ	μ	σ	μ	σ
Ours	0.862	0.025	44.0	2.98	1.47	0.031
DBoW2	N/A	N/A	15.8	3.08	4.25	0.547
AlexNet (no GRP)	2.13	0.038	405.0	17.4	80.8	0.708
AlexNet	16.6	0.658	418.0	17.8	N/A	N/A



- Online Loop Closure
- 精度召回曲线是二进制分类的一个很好的指标,但它们并不能完全证明我们的方法能够在实践中准确地闭合循环。因此,我们使用我们的模型在KITTI [44]序列00和05上的极其简单的应用来执行实时循环闭包。在该实验中,我们通过使用每第七帧进行循环检测来模拟关键帧选择。如果数据库查询分数高于先验阈值τ,则提出循环闭包假设,并且如果三个连续查询在第一查询的六个帧内检索描述符,则确定循环闭合。
- 我们从搜索空间中排除最近的图像,并且在数据库足够大之前不要启动循环检测。我们从图6(底部)中所示的精度和召回中选择τ,以便以最高精度最大化召回率。图12显示了该实验的结果。显然,我们的方法能够使用来自完全不相关的地面实况数据的阈值来一致地关闭实际SLAM数据集上的循环,这表明它已准备好在实时SLAM系统中使用。此外,我们的在线循环闭包模型的应用非常简单,并且可以通过查看k-最近邻居而不是单个最近邻居来轻松改进,添加额外的误报拒绝方法(即几何检查),或使用下一节中描述的任何方法。

OnlineLoop Closure

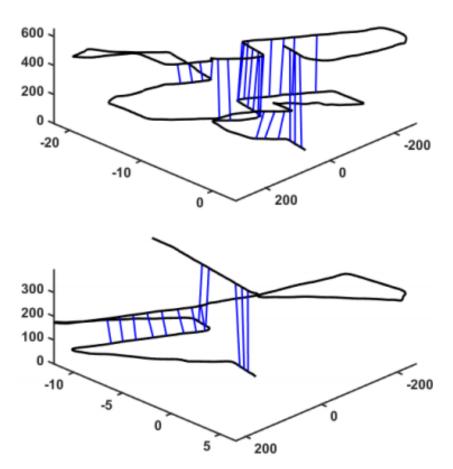
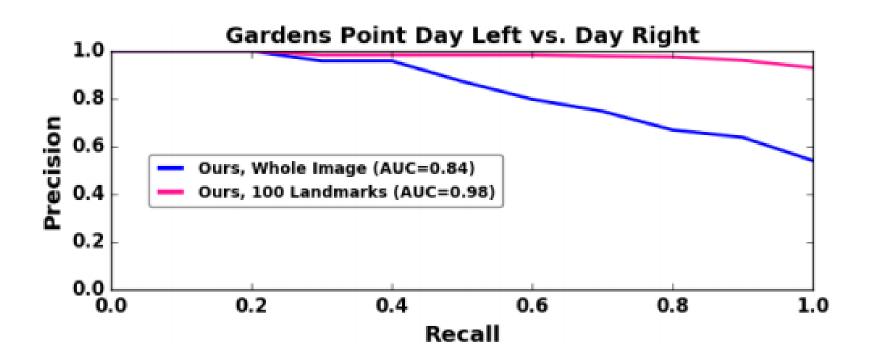


Fig. 12: The results of online loop closure using KITTI 00 and 05, respectively. The 2D location of the trajectory is

- Integration into ConvNet-Based Place Recognition Systems
- 如第III-D节所述,我们的模型可以轻松地集成到基于ConvNet的现成位置识别系统[12,32,14,33]中,以便更快地进行特征提取。这些方法建立在整体图像描述符的使用上,在不同情况下提高了性能。他们将ConvNet视为用于图像描述的黑盒子 从分类网络中丢弃图像分类。许多这些方法被迫减少ConvNet功能的尺寸以最小化运行时间,而我们的模型已经产生足够小的描述符用于实时使用,并且比这些方法使用的典型分类网络更小和更快。
- 为了证明这一点,我们重现了[12]中提出的最先进的基于地标的地点识别系统,用BING [47]替换了Edge Boxes [46]作为Hou等人。[14]确实减少了运行时间,并将尺寸减小的AlexNet conv3地标描述符替换为我们模型中的描述符。Sünderhauf等。[12]建议将AlexNet conv3层减少到1,024个维度,而我们自然是1,064个,所以我们不需要进一步减少它 避免投影m个界标描述符所需的1024×64896×64896×m矩阵乘法的成本(使用AlexNet)到1,024个维度,每次将新图像添加到数据库时都必须这样做。该实验的结果可以在图13中看到。基于地标的方法相对于整体图像描述符提供了巨大的改进 在Gardens Point白天数据集上接近完美的性能。我们的模型无缝集成到该系统中,这表明它可以轻松替换其他基于ConvNet的场所识别系统中的分类网络。



CONCLUSIONS

- 我们提出了一种新的无监督深度神经网络,用于快速和稳健的闭环,适用于视觉SLAM。基于去噪自动编码器架构,我们对图像应用随机投影变换,以捕捉由于机器人运动引起的视点的极端变化,同时采用固定长度的HOG描述符来帮助我们的网络更好地学习场景的几何。建议的模型允许在培训中使用大量数据,因为它们都不需要标记或包含任何特殊信息。此外,尽管我们的预训练模型在当前状态下概括良好,但由于我们的无监督设计,很容易进行微调或重新训练。随着更多数据的出现,增加了改进的可能性。
- 我们针对不同的数据集进行了彻底的比较研究,与最先进的地点识别图像描述方法进行了对比,其中广泛的实验结果表明,所提出的深环闭合方法通常优于基准。有效性(精确召回)和效率(运行时)。我们的模型紧凑,强大,快速 使其成为在基于ConvNet的场所识别系统中替换更大,更慢的分类网络的有希望的候选者,正如我们通过再现[12]所示。由于其轻巧而坚固的设计,我们的模型适用于实时SLAM系统 特别是直接算法[48,25,49,50],其中不需要中间图像表示。我们的目标是为循环闭合提供开箱即用的解决方案,更一般地说,是地点识别。我们目前正在努力将我们的模型集成到各种SLAM系统中,适用于具有挑战性的环境中的自主导航。

Code

- https://github.com/rpng/calc
- https://github.com/rpng/calc2.0