# 二十一组大作业报告：图像检索任务

组长：董智桥 PB14000246

组员：叶术优 PB14209100

杨宇辰 PB15061202

戴智博PB16060630

一、任务说明

我们所做的项目是图像检索任务，即从海量数据集中查找出与查询图片最相似的一张或多张图片并按照相似度大小排序返回。由于计算机科学与传感器技术的高速发展，我们生活中的数据量迅速增大，大数据时代的到来意味着如何从有用的信息中检索得到关键有效的信息成为重要的课题。而图像又包含着丰富的信息，俗话说，一图胜千言，所以图像检索技术能够挖掘从海量图像数据集中挖掘出有意义的信息。如今，图像检索技术在现实生活中有着许多应用，比如网络购物，版权保护，刑侦分析等，让我们感受颇深。正因为手机以图搜图功能和网络购物的普及，激发了我们想要探究图像检索任务的愿望。

图像检索任务普遍存在着两个难点，即语义鸿沟和维数灾难。第一个难点：语义鸿沟的含义是特征十分相似的图像可能视觉感受上完全不同。第二个难点：维数灾难的含义是由于检索数据集十分巨大，如果每张图片的特征维数较高，会带来巨大的存储代价并且严重影响检索速度，而不能达到实时检索的要求。此外，由于欧氏距离在高维空间中对于某些数据分布难以有效衡量相似性，导致太高的特征维度会十分影响检索性能。

二、数据处理

我们考虑图像检索中经典而困难的地标识别任务，采用公开数据集landmark作为训练数据集，oxford5k作为测试数据集。landmarks数据集又分为full和clean两种版本，landmark-clean包括35376张图片来自586个landmarks，landmark-full包括140372张图片来自680个landmarks。Oxford 5k数据集由5062张牛津建筑物的图片组成，来自11个著名景点，并且每个景点有五张图片作为查询图片，其余每张图片都有标签表明它是查询图片的正例样本（希望被检索出的），不清楚的正例样本（虽然是正例样本但是检索难度很大，不影响评价指标的计算）与负例样本（不希望被检索出）。Landmark-full数据集包含所有谷歌搜寻的地标建筑图片，去除了与oxford以及paris数据集重合类别的图片以免在训练时学到了测试数据集的内容。而landmark-clean数据集通过对Landmarkfull数据集的一种清洗过程，去除了虽然属于同一类但是视觉上不相似的所有图片。

在训练过程中，我们采用数据增强技术。对于我们模型的第一阶段训练，我们将输入图片先剪切成为方形，然后调整尺度到250\*250，然后随机剪切为224\*224的图片作为输入，对于我们模型的第二阶段训练，我们先将图片剪切为方形，然后调整尺寸到900\*900，再随机剪切到720\*720，最后随机乘以一个尺度因子（小于等于1）再输入到网络中。尺度因子小于等于1的原因是因为最后提取的局部特征数目与输入图片的实际大小有关，所以尺度因子超过1的话容易导致gpu显存溢出而训练失败。

在我们采用分类损失进行训练时，将图片与其所属的建筑物类别一起输入到网络中，最后类别的数目即建筑物类别的数量。在我们采用对比损失/三元组损失进行训练时，我们需要挖掘相似/不相似的图片组成正例/负例样本对或（正，正，负）三元组，此时，我们采用landmark-clean数据集，由于其同一类的图片视觉上相似，所以我们认为同一类的图片互相可以作为正例样本，而不同类的图片便可以作为负例样本。

三、算法原理

我们采用的算法是Delf（Deep local feature），它是一种深度局部特征。相比于图像检索任务中常用的基于SIFT的局部特征和基于CNN的全局特征，这种方法一定程度上结合了两者的优点而避免了两者的缺点。这是由于基于SIFT的局部特征是根据海森探测器得到每张图片中具有尺度不变性的关键点描述符，然后计算关键点附近的梯度从而为每张图片生成的大量局部特征，这种传统的方法可以有效的检测每张图片中边缘、角点等特征，但是难以跨越语义鸿沟，会导致虽然图片特征相似而视觉上差别很大。而基于CNN的全局特征由于卷积核对于变换（平移，旋转）的敏感性，导致基于CNN的全局特征也对诸如光照变化、视觉变化、旋转等变换的鲁棒性较差。而深度局部特征是一种基于CNN网络提取得到的局部特征，由于深度神经网络的特性可以较好的跨越语义鸿沟，而由于局部特征后续要经过聚类操作，可以一定程度上增加鲁棒性，较好的解决其他方法存在的问题。

Delf（深度局部特征）的提取分为三个步骤，即稠密局部特征提取，基于注意力机制的关键点选择和多尺度紧凑特征提取。前两个步骤是模型的训练步骤，第三个步骤是特征的提取。本文采用Resnet50模型，使用其第三或第四层卷积层后输出的特征图作为提取的稠密特征。第一步即在landmark-clean数据中通过分类损失对模型再次训练以进行微调，这是因为实验证明在与目标数据集（oxford5k）分布更为相似的数据集上微调网络模型可以提高模型的特征表达能力。然而直接使用稠密的特征图作为局部特征会导致每张图像的局部特征数量太大，产生非常多的冗余信息。因为这些局部特征是遍历整幅图像生成的，所以会将诸如背景，行人等遮挡与噪音的特征引入，从而而使得真正对检索有意义的局部特征变得稀疏。第二步即将第一步得到的稠密特征（特征图）通过一个两层CNN（注意力机制）来为每个局部特征生成一个得分，再将这些局部特征加权求和后输入进全连接层以分类损失进行训练，这一步的训练数据集采用landmark-full，类别的数量即建筑物类别的数量。在训练好网络模型后，由于提取得到的稠密局部特征都来自Resnet50的同一层，即所有局部特征对于原图的感受野都是固定大小的。而一个很本能的想法是我们希望图像的局部特征中能包含图像不同尺度的细节信息，所以在测试的时对输入的每张查询图片调整到七个尺度上，即0.25，0.3535，0.5，0.71，1.0，1.4141，2.0。然后在提取的所有局部特征中，选择1000个注意力得分最高的，作为每张图片的特征表达。然后对这些特征进行l2归一化再进行PCA降维到40维，再重新进行一次l2归一化，最终对每张图片得到1000个40维紧致且具有辨别力的特征。在得到局部特征后，我们采用经典的VLAD聚类算法对其进行聚类生成全局特征，然后通过特征点积计算图片间的相似度并进行排序。

四、实验细节

在提出的模型训练过程中，特征描述符和注意力模型都是隐式的只通过图片级别的标签学习得到的。然而这给学习过程带来了一些挑战，因为图像的特征表达和特征的评价函数会一起通过反向传播而训练，我们发现这样训练实际的效果很差，由其是评价函数无法正确的收敛而会产生随机性很强的波动。因此，我们采用了一种两阶段的训练策略，即我们先通过微调模型学习图像的特征表达而不加以训练评价函数，然后再单独训练评价函数使用固定的特征表达而不再训练特征提取部分。

我们首先在landmark-clean数据集上微调提取DELF特征的ResNet50模型，然后采用landmark-full数据集训练提取DELF特征的关键点选择部分。两次训练的学习率都为0.008，批次大小为8。在VLAD聚类算法中，经过对k值的多次尝试，设定k=256是一个兼具性能与检索效率的选择。

我们尝试了采用网络的第三层或第四层作为稠密的特征表达，这是因为不同深度神经网络的泛化能力不同，而由于受特征图大小限制，不能选择太小或太大的特征图。在第一步微调Resnet50模型时我们尝试了将网络改为Siamese网络，并采用对比损失/三元组损失进行训练，正例与负例样本的挖掘如第二节数据处理所述，我们尝试了对比损失与三元组损失中不同的margin如0，0.1，0.5等，但效果皆不佳，不如采用分类损失微调模型的效果好。我们认为这是因为挖掘正例与负例样本对时无法完全利用到数据集的信息，会使得模型出现一定的偏向性，而影响检索性能。最终训练的epoch数目第一阶段和第二阶段皆为30。在oxford数据集上以map为评价指标测试的结果如下：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| K | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |
| DELF+Vlad | 51.58 | 51.51 | 57.28 | 63.07 | 65.65 | 65.25 | 60.2 |

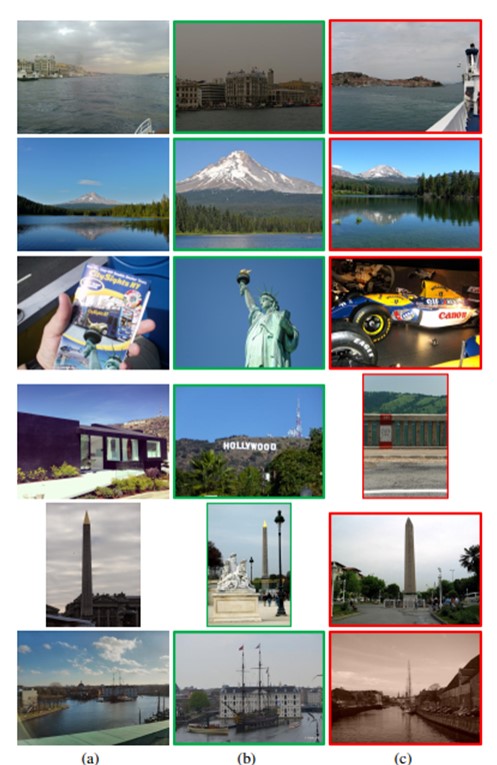
五、实验总结

在本次项目实践中，我们通过对神经网络结构的调整，学习率，批次大小以及各种参数的调整，清楚的感受到了神经网络前向传播和bp的过程，了解了神经网络强大的学习能力与迁移泛化能力。

我们尝试了多种学习率，如0.1，0.01，0.001，0.005等，我们发现学习率设置过大会导致网络无法训练，出现nan的情况，而学习率设置太小会导致网络最后达到的性能较差，我们认为这是网络收敛速度太慢，经过多次迭代后梯度消失的原因，所以设置合适的学习率大小至关重要。此外，在设置批次大小时，较大的批次可以加快网络的收敛速度，且显著提高训练速度，我们认为这是由于批次较大可以提高网络由局部最优调整为全局最优的速度，而不会困于局部最优的鞍点。不过过大的批次会导致gpu显存溢出，所以应在合适的范围内调大批次。在设置优化方式时，我们尝试了SGD与ADAM，并没有明显的区别，学习率的迭代采用每k个epoc乘以0.5的方式，尝试了k=1，5，10，其中k=10的收敛速度最快。

我们通过设置不同的训练集与训练过程，也清晰的了解了训练过程对网络的意义非凡。在采用三元组损失训练网络时，由于数据集挖掘的不够充分使得模型出现了一定的偏向性导致性能下降。在一起训练特征提取部分与注意力机制时，注意力机制部分网络完全不能收敛，表现出很强的随机性，而换成二阶段训练策略时有了明显改善，这告诉我们设计合理的训练过程意义重大。在将图片输入进网络提取特征时，由于数据集中某些特殊图片会导致返回的特征为空，这造成了我们很大的困扰，在代码中加入合适的错误处理与错误检查后才避免了这种情况，而由于出错的情况很少，错误处理也不会对性能有很大影响。

对实验性能的分析以及可视化效果的分析中，我们清楚的感受到了深度局部特征优秀的检索性能。如下图



六、分工与合作情况

戴智博：DELF模型训练部分学习，修改以及分析

董智桥：DELF模型提取部分学习，修改以及分析

叶术优：基于DELF局部特征进行聚类与检索

杨宇辰：想法提出与数据集处理