Unbiased Metric Learning: On the Utilization of Multiple Datasets and Web Images for Softening Bias

无偏度度量学习: 利用多个数据集和网络图像来减轻偏差

Chen Fang Ye Xu Daniel N. Rockmore
陈方叶许丹尼尔 · N · 罗克莫尔
Computer Science Department
计算机科学系
Dartmouth College
达特茅斯学院
Hanover, NH 03755, U.S.A.
美国新罕布什尔州汉诺威 03755
{chenfang, ye, rockmore}@cs.dartmouth.edu
{chenfang, ye, rockmore}@cs.dartmouth.edu

Abstract

摘要

Many standard computer vision datasets exhibit biases due to a variety of sources including illumination condition, imaging system, and preference of dataset collectors. Biases like these can have downstream effects in the use of vision datasets in the construction of generalizable techniques, especially for the goal of the creation of a classification system capable of generalizing to unseen and novel datasets. In this work we propose Unbiased Metric Learning (UML), a metric learning approach, to achieve this goal. UML operates in the following two steps: (1) By varying hyperparameters, it learns a set of less biased candidate distance metrics on training examples from multiple biased datasets. The key idea is to learn a neighborhood for each example, which consists of not only examples of the same category from the same dataset, but those from other datasets. The learning framework is based on structural SVM. (2) We do model validation on a set of weakly-labeled web images retrieved by issuing class labels as keywords to search engine. The metric with best validation performance is selected. Although the web images sometimes have noisy labels, they often tend to be less biased, which makes them suitable for the validation set in our task. Cross-dataset image classification experiments are carried out. Results show significant performance improvement on four well-known computer vision datasets.

许多标准计算机视觉数据集由于多种来源 (包括照明条件、成像系统和数据集收集者的偏好) 而表现出偏差。这些偏差可能对视觉数据集在构建可推广技术中的使用产生下游影响，特别是在创建能够推广到未见和新颖数据集的分类系统的目标上。在本工作中，我们提出了无偏度度量学习 (UML)，这是一种度量学习方法，以实现这一目标。UML 通过以下两个步骤进行操作:(1) 通过变化超参数，它在来自多个偏见数据集的训练示例上学习一组偏差较小的候选距离度量。关键思想是为每个示例学习一个邻域，该邻域不仅包括来自同一数据集的同一类别的示例，还包括来自其他数据集的示例。学习框架基于结构化支持向量机。(2) 我们对通过将类别标签作为关键字发给搜索引擎检索到的一组弱标记网络图像进行模型验证。选择验证性能最佳的度量。尽管网络图像有时具有噪声标签，但它们往往倾向于较少偏见，这使得它们适合用于我们任务中的验证集。进行了跨数据集图像分类实验。结果显示在四个著名计算机视觉数据集上显著提高了性能。

1. Introduction

1. 引言

Over the last decades object recognition systems have been improved dramatically [26][19][6]. One of the forces driving the development is the availability of medium or large size high quality image datasets (e.g., Caltech 101 [7], PASCAL VOC [4], LabelMe [20] and SUN09 [3]) that enable researchers to evaluate and practice sophisticated feature designing and machine learning techniques. However,

在过去几十年中，物体识别系统得到了显著改善 [26][19][6]。推动这一发展的力量之一是中等或大型高质量图像数据集的可用性 (例如，Caltech 101 [7]、PASCAL VOC [4]、LabelMe [20] 和 SUN09 [3])，这些数据集使研究人员能够评估和实践复杂的特征设计和机器学习技术。然而，

Torralba and Efros [23] point out that every dataset carries bias in its own way, which can be caused by various reasons, such as illumination condition, different imaging system and preference of database collectors. Bias inevitably leads learning algorithms to overfit for the training set to the detriment of any ability to generalize to other datasets. In other words, an object recognition system trained solely on one dataset tends to perform poorly on unseen and novel datasets at test time, because the underlying bias is incorporated into the learning algorithm. This is an important issue, because most image classification systems are required to handle all kinds of test examples, regardless of where the test examples are drawn from. For example, it's rare to see a question like "Can you classify a dog image from Caltech101?".

Torralba 和 Efros [23] 指出，每个数据集以其自身的方式携带偏差，这可能由多种原因造成，例如照明条件、不同的成像系统和数据库收集者的偏好。偏差不可避免地导致学习算法对训练集过拟合，从而损害其对其他数据集的泛化能力。换句话说，单独在一个数据集上训练的物体识别系统在测试时往往在未见过的新数据集上表现不佳，因为潜在的偏差被纳入了学习算法。这是一个重要问题，因为大多数图像分类系统需要处理各种测试示例，无论这些测试示例来自何处。例如，很少会看到“你能对来自 Caltech101 的狗图像进行分类吗?”这样的问题。

The new challenge now is to build a system that performs well on unseen datasets. This requires the learned model to capture the general class knowledge, while discarding the information reflective of the bias. Note that this problem is potentially more general than a challenge facing researchers performing image analysis. Microarray analysis is one context in which this issue has received significant attention (see e.g., [1]).

现在的新挑战是构建一个在未见过的数据集上表现良好的系统。这要求学习到的模型能够捕捉一般类别知识，同时丢弃反映偏差的信息。请注意，这个问题可能比面临图像分析的研究人员所面临的挑战更为普遍。微阵列分析就是一个在这个问题上受到显著关注的背景 (见例如 [1])。

Failing the ability to create a vision dataset free of bias, it is of interest to address the assumption of bias directly. To this end we introduce Unbiased Metric Learning (UML), which learns an unbiased metric using multiple biased datasets and web images. Our approach operates in the following two steps:

在无法创建一个没有偏差的视觉数据集的情况下，直接解决偏差的假设是有意义的。为此，我们引入了无偏度度量学习 (UML)，它利用多个有偏数据集和网络图像学习无偏度度量。我们的方法分为以下两个步骤:

Step 1 - we learn a set of distance metrics on training examples from multiple biased datasets. The key idea is that in the learned feature spaces, the neighborhoods of training examples should consist of not only examples of the same category from the same dataset, but those examples of the same category from other datasets. We call this property "neighborhood diversity." In such a way, datasets (or domains) are bridged together via these neighborhoods. By varying related hyperparameters, we learn a set of distance metrics, each of which bears a specific degree of "neighborhood diversity." In other words, the training data is distributed differently in the spaces defined by these metrics. We form a candidate set of these metrics for Step 2 (below), in which a novel validation is performed. Technically, we cast it as a learning to rank problem [11][27], and solve it with structural metric learning [15].

第一步 - 我们从多个有偏数据集中学习一组距离度量的训练示例。关键思想是在学习到的特征空间中，训练示例的邻域不仅应由来自同一数据集的同类示例组成，还应包括来自其他数据集的同类示例。我们称这种特性为“邻域多样性”。通过这种方式，数据集 (或领域) 通过这些邻域相互连接。通过改变相关的超参数，我们学习一组距离度量，每个度量具有特定程度的“邻域多样性”。换句话说，训练数据在这些度量定义的空间中分布不同。我们为第二步 (如下) 形成一组候选度量，其中进行了一种新颖的验证。从技术上讲，我们将其视为一个学习排序问题 [11][27]，并通过结构度量学习 [15] 来解决。

Step 2 - we do model validation to identify the metric with best generalization ability. Conventionally, the validation set is from the same source as the training set, for example, cross-validation. However, in our case, a good validation performance on seen datasets does not necessarily generalize to unseen datasets. Therefore conventional validation may fail to uncover the desired model. So, instead of using images from seen datasets, we propose to use a set of weakly-labeled web images retrieved from the Internet by issuing class labels as keywords to the search engine. Those images are less biased and have higher intra-class variability than human collected datasets, which makes it suitable to be used as our validation set.

第二步 - 我们进行模型验证，以识别具有最佳泛化能力的度量。传统上，验证集来自与训练集相同的来源，例如交叉验证。然而，在我们的案例中，在已见数据集上的良好验证性能并不一定能推广到未见数据集。因此，传统验证可能无法发现所需的模型。因此，我们建议使用一组从互联网上检索的弱标记网页图像，作为关键字向搜索引擎发出类别标签。这些图像的偏差较小，类内变异性高于人类收集的数据集，使其适合作为我们的验证集。

We do image classification experiments, and use cross-dataset performance to measure a model's ability to generalize to unseen datasets. Experiments show the following facts: (1) In Step 1, by varying

hyperparameters, our learning framework is capable of producing models with superior cross-dataset classification performance. (2) In the validation step, the model selected by our novel validation method significantly outperforms those selected by conventional validation procedures.

我们进行图像分类实验，并使用跨数据集性能来衡量模型对未见数据集的泛化能力。实验显示以下事实:(1) 在第一步中，通过改变超参数，我们的学习框架能够生成具有优越跨数据集分类性能的模型。(2) 在验证步骤中，采用我们新颖验证方法选择的模型显著优于那些通过传统验证程序选择的模型。

In the rest of the paper, we will first review related work, discuss the advantages of our framework, then cover technical details. We then follow this with experiments demonstrating the effectiveness of our approach.

在本文的其余部分，我们将首先回顾相关工作，讨论我们框架的优势，然后涵盖技术细节。接着，我们将进行实验，展示我们方法的有效性。

2. Related Work

2. 相关工作

The issue of dataset bias in vision datasets was first raised by Torralba and Efros [23]. Since then, Khosla et al. [12] has proposed a solution to improve cross-dataset generalization ability for an object recognition system. The approach in [12] learns a common weight vector which is expected to work well on unseen datasets, in a way that is similar to regularized multi-task learning [5]. Our method adopts a metric learning solution to this problem.

视觉数据集中数据集偏差的问题最早由 Torralba 和 Efros 提出 [23]。此后，Khosla 等人 [12] 提出了一个解决方案，以提高物体识别系统的跨数据集泛化能力。[12] 中的方法学习一个共同的权重向量，期望其在未见数据集上表现良好，这种方式类似于正则化的多任务学习 [5]。我们的方法采用了一种度量学习的解决方案来应对这个问题。

Metric learning methods have been popular in the machine learning community as well as computer vision [25][18]. Most of them can be categorized as either learning a "global" metric or a "local" metric. Our approach is closer to the former category, where a single parametric transformation is learned to map data from the original space to a new space, where the data distribution exhibits some desired properties. Weinberger et al. [24] propose a large margin method to group together examples with the same label and separate the ones with different labels. In [9] a metric is learned by collapsing all examples in a class to a single point. However, these methods are not designed to utilize domain information. As shown in later experiments, this results in poor cross-dataset generalization performance.

度量学习方法在机器学习社区以及计算机视觉领域都很受欢迎 [25][18]。它们大多数可以分为学习 "全局" 度量或 "局部" 度量。我们的方法更接近前者，其中学习一个单一的参数化变换，将数据从原始空间映射到一个新空间，在这个新空间中，数据分布展现出一些期望的特性。Weinberger 等人 [24] 提出了一个大间隔方法，将具有相同标签的示例聚集在一起，并将不同标签的示例分开。在 [9] 中，通过将一个类别中的所有示例压缩到一个单一的点来学习度量。然而，这些方法并未设计用于利用领域信息。如后续实验所示，这导致了较差的跨数据集泛化性能。

Because our approach utilizes domain information in training data, it is related to domain adaptation and multitask learning. In domain adaptation, the goal is to transfer knowledge from the source domain to help perform a task in the target domain, and multi-task learning aims at good performance simultaneously in multiple domains. Among the large range of work in this area, [21] and [17] are both metric learning-based. [21] learns a metric to transfer knowledge to the target domain by randomly sampling pairs consisting of a labeled example from the source domain and another from the target domain, and constraining their distance to be no greater (less) than a bound if the labels are the same (different). In [17], the metric is decomposed into a domain specific part and a global part, which is shared among all domains. Recent work by Tommasi et al. [22] proposes to learn general knowledge from multiple datasets, which will be transferred to a target test dataset later with the help of a few labeled examples from the target dataset. Although related, our problem is different. In our problem, there is no target dataset where the task will be performed, not to mention labeled examples to assist in knowledge transfer. This means that incoming test examples may come from any domains, although most are unseen at training time. Therefore, the goal is to learn a transformed new space, where all the training examples are less biased. So any classification system trained on it will generalize better to novel datasets.

因为我们的方法在训练数据中利用了领域信息，所以它与领域适应和多任务学习相关。在领域适应中，目标是将知识从源领域转移，以帮助在目标领域执行任务，而多任务学习旨在在多个领域中同时实现良好的性能。在这一领域的大量研究中，[21] 和 [17] 都是基于度量学习的。[21] 通过随机抽样由源领域中

的标记示例和目标领域中的另一个示例组成的对，学习一个度量，以将知识转移到目标领域，并约束它们的距离在标签相同 (不同) 时不大于 (小于) 一个界限。在 [17] 中，度量被分解为一个领域特定部分和一个在所有领域中共享的全局部分。Tommasi 等人最近的工作 [22] 提出了从多个数据集中学习一般知识的方案，这些知识将在稍后通过目标数据集中的少量标记示例转移到目标测试数据集中。尽管相关，我们的问题是不同的。在我们的问题中，没有目标数据集可以执行任务，更不用说用于辅助知识转移的标记示例。这意味着即将到来的测试示例可能来自任何领域，尽管大多数在训练时是未见的。因此，目标是学习一个转化的新空间，使所有训练示例的偏差更小。因此，任何在该空间上训练的分类系统将更好地推广到新数据集。

3. Approach Overview

3. 方法概述

Our approach tackles the problem from a learning-to-rank perspective. First, we consider each training example as a query and rank in ascending order the other training examples based on their L^2 distances to the query. We want a high precision among top- k positions. This is essentially constructing a label-coherent neighborhood. Then, beyond the neighborhood label coherence, we look at the domain information of the positive examples within the neighborhood. We encourage to include more positive examples from multiple domains, which was mentioned before as "neighborhood diversity." From a ranking perspective, this is inserting into the top- k positions, those positive examples from domains other than the current query's. Fig. 1 gives a schematic comparison of the case in which "neighborhood diversity" is enforced and the case in which it is not. As the top right figure illustrates, examples of the same category from different datasets are linked together via diversified neighborhood, which is constructed automatically in the learning procedure by discovering and grouping appropriate pairs of positive examples from different datasets. However, we still let each dataset keep its own distributional independence, since we do not want to overly "merge" the data. We let the hyperparameters vary, so as to produce a set of metrics with different levels of "neighborhood diversity" which later will be screened by validation procedure.

我们的方法从学习排序的角度解决这个问题。首先，我们将每个训练示例视为一个查询，并根据它们与查询的 L^2 距离按升序对其他训练示例进行排序。我们希望在前 k 个位置中达到高精度。这本质上是在构建一个标签一致的邻域。然后，超越邻域标签一致性，我们关注邻域内正例的领域信息。我们鼓励从多个领域中包含更多正例，这在之前被称为 "邻域多样性"。从排名的角度来看，这就是将来自当前查询以外领域的正例插入到前 k 个位置。图 1 给出了强制执行 "邻域多样性" 的情况与不强制执行的情况的示意比较。如右上角的图所示，来自不同数据集的同一类别的示例通过多样化的邻域连接在一起，这在学习过程中通过发现和分组来自不同数据集的适当正例对自动构建。然而，我们仍然让每个数据集保持其自身的分布独立性，因为我们不想过度 "合并" 数据。我们让超参数变化，以产生一组具有不同 "邻域多样性" 水平的度量，这些度量随后将通过验证程序进行筛选。

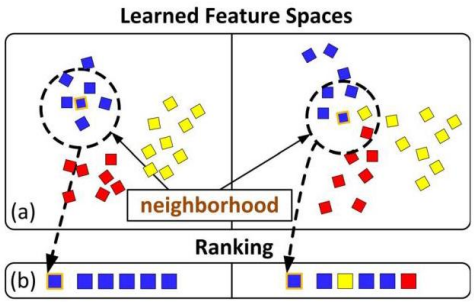

Figure 1. Schematic illustration of neighborhoods without (top-left) and with (top-right) "neighborhood diversity" (best viewed in color). (a) The distribution of data in different learned feature spaces (Left: a normal feature space. Right: a feature space learned by UML). Color and shape represent dataset and category, respectively. Squares with colored boundaries are the corresponding queries of the two local neighborhoods in circle. In the top right figure, data from different datasets are linked at circled neighborhoods. (b) A look at the local neighbors in Fig.1a from a ranking perspective. The ranking on the right, unlike the one on the left, exhibits both label coherence and neighborhood diversity.

图 1. 没有 (左上) 和有 (右上) “邻域多样性” 的邻域示意图 (最佳在彩色下查看)。(a) 不同学习特征空间中数据的分布 (左: 正常特征空间。右: 由 UML 学习的特征空间)。颜色和形状分别代表数据集和类别。带有彩色边界的方形是圆中两个局部邻域的相应查询。在右上图中，来自不同数据集的数据在圈定的邻域中相连。(b) 从排名的角度看图 1a 中的局部邻居。右侧的排名与左侧不同，展现了标签一致性和邻域多样性。

As for model validation, we use weakly-labeled web images as the validation set. Images returned by search engines are sometimes loosely related to textual query (keyword), compared to human-labeled datasets. Therefore, if used as training set, in order to get comparable results, extra labor is needed to remove noise and build a robust learning algorithm [8][10]. However, studies have shown that they serve well as a relatively noisy validation set [8]. In our case, we are interested in the fact that web images tend to be less biased, since they are implicitly sampled from a countless number of unknown sources and there is less human intervention during collection process. Our experiments support the claim that web images serve better as a validation set for our problem.

关于模型验证，我们使用弱标记的网络图像作为验证集。与人工标记的数据集相比，搜索引擎返回的图像有时与文本查询 (关键词) 关系较松散。因此，如果用作训练集，为了获得可比的结果，需要额外的劳动来去除噪声并构建一个稳健的学习算法 [8][10]。然而，研究表明，它们作为相对嘈杂的验证集表现良好 [8]。在我们的案例中，我们关注的是网络图像往往较少偏见，因为它们是从无数未知来源隐式采样的，并且在收集过程中人类干预较少。我们的实验支持网络图像作为我们问题的验证集更为有效的说法。

4. Technical Details

4. 技术细节

In this section, we will discuss the formulation of the structural metric learning (SML) framework as well as Un - biased Metric Learning (UML). Technical details will be covered.

在本节中，我们将讨论结构度量学习 (SML) 框架的公式以及 Un - 偏置度量学习 (UML)。将涵盖技术细节。

4.1. Notations and preliminaries

4.1. 符号和预备知识

Let $\mathcal{D} = \{D_1, \ldots, D_Q\}$ denote the set of biased datasets that share a set of common classes $\mathcal{C} = \{c_1, \ldots, c_K\}$. The training set is denoted as $\mathcal{X}$ with $|\mathcal{X}| = n$. Each training example $i \in \mathcal{X}$ is a triplet (x_i, l_i, d_i) , where $x_i \in \mathbb{R}^d$ is the feature vector of item $i, l_i \in \mathcal{C}$ is the corresponding class label and $d_i \in \mathcal{D}$ indicates the dataset that i is from. For a query q , we use $\mathcal{X}_q^+ / \mathcal{X}_q^-$ to denote the set of positive/negative examples in $\mathcal{X}$.$\mathcal{Y}$ will be the set of permutations/rankings of items in $\mathcal{X}$. Similarly, $\mathcal{Y}_q^+$ is the set of permutations/rankings of items in $\mathcal{X}_q^+$. If i is ranked before (after) j in some ranking $y \in \mathcal{Y}$, we say $i \prec_y j$ $(i \succ_y j)$.

令 $\mathcal{D} = \{D_1, \ldots, D_Q\}$ 表示共享一组公共类别 $\mathcal{C} = \{c_1, \ldots, c_K\}$ 的偏置数据集的集合。训练集表示为 $\mathcal{X}$ ，包含 $|\mathcal{X}| = n$ 。每个训练示例 $i \in \mathcal{X}$ 是一个三元组 (x_i, l_i, d_i) ，其中 $x_i \in \mathbb{R}^d$ 是项目的特征向量，$i, l_i \in \mathcal{C}$ 是相应的类别标签，而 $d_i \in \mathcal{D}$ 表示 i 来自的数据集。对于查询 q ，我们用 $\mathcal{X}_q^+ / \mathcal{X}_q^-$ 表示在 $\mathcal{X}.\mathcal{Y}$ 中的正/负示例集合，而 $\mathcal{X}.\mathcal{Y}$ 将是 $\mathcal{X}$ 中项目的排列/排名集合。类似地，$\mathcal{Y}_q^+$ 是 $\mathcal{X}_q^+$ 中项目的排列/排名集合。如果 i 在某个排名 $y \in \mathcal{Y}$ 中排在 j 之前 (之后)，我们称之为 $i \prec_y j$ $(i \succ_y j)$ 。

For a matrix $W \subset \mathbb{R}^{d \times d}, W \succcurlyeq 0$ means it is a symmetric and positive semidefinite matrix. The Mahalanobis distance defined by W is $d_{\mathbf{W}}(i,j) = \sqrt{(x_i - x_j)^\top \mathbf{W} (x_i - x_j)}$. This is equivalent to applying a transformation $L \subset \mathbb{R}^{d' \times d}$ to the original feature space and calculating the L^2 distance in the new space. Thus, $W = L^\top \times L$. The Frobenius inner product of two matrices $A, B \in \mathbb{R}^{d \times d}$ is denoted as $\langle A, B \rangle_F = \operatorname{tr}\left(A^\top B\right)$. Finally, $\mathbf{1}(\mathbf{X})$ is the indicator function of event $\mathbf{X}$.

矩阵 $W \subset \mathbb{R}^{d \times d}, W \succcurlyeq 0$ 意味着它是一个对称的半正定矩阵。由 W 定义的马哈拉诺比斯距离为 $d_{\mathbf{W}}(i,j) = \sqrt{(x_i - x_j)^\top \mathbf{W} (x_i - x_j)}$ 。这等同于对原始特征空间应用一个变换 $L \subset \mathbb{R}^{d' \times d}$ ，并在新空间中计算 L^2 距离。因此，$W = L^\top \times L$ 。两个矩阵的弗罗贝纽斯内积 $A, B \in \mathbb{R}^{d \times d}$ 表示为 $\langle A, B \rangle_F = \operatorname{tr}\left(A^\top B\right)$ 。最后，$\mathbf{1}(\mathbf{X})$ 是事件 $\mathbf{X}$ 的指示函数。

4.2. Structural metric learning

4.2. 结构度量学习

We now review the structural metric learning (SML) framework [15]. The goal is to learn a positive semidefinite matrix W , so that when a query q is issued, the corresponding ranking or ordering y_q , which is produced based on the Mahalanobis distance defined by W , will have some desiring properties, such as high Precision@k, Mean Average Precision or ROC area. This can be solved via structural learning and its mathematical formulation is similar to structural SVM [11]. The following objective function is minimized:

我们现在回顾结构度量学习 (SML) 框架 [15]。其目标是学习一个正半定矩阵 W ，以便在发出查询 q 时，基于 W 定义的马哈拉诺比斯距离产生的相应排名或排序 y_q 将具有一些期望的属性，例如高 Precision@k、平均平均精度或 ROC 面积。这可以通过结构学习来解决，其数学公式与结构支持向量机 (SVM) 类似 [11]。以下目标函数被最小化:

$$\min_{W\succcurlyeq 0,\xi\geq 0} \operatorname{tr}(W) + \frac{C}{n}\sum_{q\in\mathcal{X}} \xi_q \tag{1}$$

subject to constraints as follows:

受以下约束条件限制:

$$\forall q \in \mathcal{X}, \forall y \in \mathcal{Y} \setminus y_q^* : \tag{2}$$

$$\left\langle W, \psi_{po}\left(q, y_q^*\right) - \psi_{po}(q, y)\right\rangle_F \geq \Delta\left(y, y_q^*\right) - \xi_q.$$

in Eq. 1$\operatorname{tr}(W)$ is the regularizer. In Eq. 2, y_q^* is the ground truth ranking for query q . The right-hand side of Eq. 2 includes the slack variable ξ_q and the structural loss function $\Delta\left(y, y_q^*\right)$, which encodes the structural information in $\mathcal{Y}$. Thus, the margin is rescaled to be $\Delta\left(y, y_q^*\right)$. The Frobenius inner product term on the left is the discriminative score difference between current ranking y and y_q^* . $\psi(q, y)$ is the partial order feature [11] with the following form:

在方程 1$\operatorname{tr}(W)$ 中是正则化项。在方程 2 中，y_q^* 是查询 q 的真实排名。方程 2 的右侧包括松弛变量 ξ_q 和结构损失函数 $\Delta\left(y, y_q^*\right)$ ，它编码了 $\mathcal{Y}$ 中的结构信息。因此，边际被重新缩放为 $\Delta\left(y, y_q^*\right)$ 。左侧的弗罗贝尼乌斯内积项是当前排名 y 和 y_q^* 之间的判别得分差异。$\psi(q, y)$ 是具有以下形式的偏序特征 [11]:

$$\psi_{po}(q, y) = \sum_{i\in\mathcal{X}_q^+}\sum_{j\in\mathcal{X}_q^-} y_{ij}\left(\frac{\phi(q, i) - \phi(q, j)}{\left|\mathcal{X}_q^+\right| \cdot \left|\mathcal{X}_q^-\right|}\right) \tag{3}$$

where

其中

$$y_{ij} = \begin{cases} +1 & i \prec_y j \\ -1 & i \succ_y j \end{cases} \tag{4}$$

$\phi(q, i)$ is a feature map that captures how q and i are related. Thus, a ranking y for query q is encoded by $\psi_{po}(q, y)$ in a feature space by examining all possible relevant/irrelevant pairs. At prediction time, given a fixed W , the ranking y that maximizes the discriminative score $\langle W, \psi_{po}(q, y)\rangle_F$ is simply the collection of $i \in \mathcal{X}$ sorted by descending $\langle W, \phi(q, i)\rangle_F$. Now choose ϕ to be $\phi(q, i) = -(x_q - x_i)(x_q - x_i)^\top$. Because $d_{\mathbf{W}}^2(q, i) = \langle W, (x_q - x_i)(x_q - x_i)^\top\rangle_F$, we see that the y that maximizes $\langle W, \psi_{po}(q, y)\rangle_F$ is simply $i \in \mathcal{X}$ sorted by ascending $d_{\mathbf{W}}^2(q, i)$.

$\phi(q, i)$ 是一个特征图，捕捉了 q 和 i 之间的关系。因此，查询 q 的排名 y 通过检查所有可能的相关/无关对在特征空间中由 $\psi_{po}(q, y)$ 编码。在预测时，给定一个固定的 W ，最大化区分评分 $\langle W, \psi_{po}(q, y)\rangle_F$ 的排名 y 只是按降序排列的 $i \in \mathcal{X}$ 的集合。现在选择 ϕ 为 $\phi(q, i) = -(x_q - x_i)(x_q - x_i)^\top$ 。因为 $d_{\mathbf{W}}^2(q, i) = \langle W, (x_q - x_i)(x_q - x_i)^\top\rangle_F$ ，我们看到最大化 $\langle W, \psi_{po}(q, y)\rangle_F$ 的 y 只是按升序排列的 $i \in \mathcal{X}$ 。

4.3. Unbiased Metric Learning

4.3. 无偏度度量学习

Learning model Our model extends the original SML. First, note that the neighborhood label coherence is already forced by Eq. 2, if we set the loss function Δ to be the following form:

学习模型我们的模型扩展了原始的 SML。首先，注意到邻域标签一致性已经通过公式 2 强制执行，如果我们将损失函数 Δ 设置为以下形式:

$$\Delta\left(y, y_q^*\right) = 1 - \operatorname{Prec}@k\left(y\right) \tag{5}$$

where $\operatorname{Prec}@k\left(y\right)$ gives the percentage of positive examples among the top- k positions of ranking y , which is equivalent to the notion of label coherence in a neighborhood. To incorporate the "neighborhood diversity" property, the following constraint is added to the SML formulation:

其中 $\operatorname{Prec}@k\left(y\right)$ 表示排名 y 的前 k 个位置中正例的百分比，这等同于邻域中的标签一致性概念。为了结合 "邻域多样性" 特性，以下约束被添加到 SML 公式中:

$$\forall q \in \mathcal{X}, \forall y^+ \in \mathcal{Y}^+ \setminus y_q^{+*}: \tag{6}$$

$$\left\langle W, \psi_{po}\left(q, y_q^{+*}\right) - \psi_{po}\left(q, y^+\right)\right\rangle_F \geq \widehat{\Delta}\left(y^+, y_q^{+*}\right) - \xi_q^+$$

Assume the current query is $q = (x_q, l_q, d_q)$. In Eq. 6, $y^+ \in \mathcal{Y}_q^+$ is a subset ranking of only $i \in \mathcal{X}_q^+$. We let y_q^{+*} be the ground truth ranking in $\mathcal{Y}_q^+$ with the following property:

假设当前查询为 $q = (x_q, l_q, d_q)$ 。在公式 6 中，$y^+ \in \mathcal{Y}_q^+$ 是仅包含 $i \in \mathcal{X}_q^+$ 的子集排名。我们让 y_q^{+*} 为 $\mathcal{Y}_q^+$ 中的真实排名，具有以下属性:

$$\forall i, j \in \mathcal{X}_q^+ : i \prec_{y_q^{+*}} j, \text{ if } \left(d_i \neq d_q \wedge d_j = d_q\right) \tag{7}$$

where d_i indicates the dataset containing example i . Eq. 7 means the all the positive examples not in d_q precede those in d_q . We also have a new objective function by adding the slack variable ξ_q^+ to Eq.1:

其中 d_i 表示包含示例 i 的数据集。公式 7 意味着所有不在 d_q 中的正例都优先于那些在 d_q 中的正例。我们还通过将松弛变量 ξ_q^+ 添加到公式 1 中获得了一个新的目标函数:

$$\min_{W \succcurlyeq 0, \xi \geq 0} \operatorname{tr}\left(W\right) + \frac{C_1}{n} \sum_{q \in \mathcal{X}} \xi_q + \frac{C_2}{n} \sum_{q \in \mathcal{X}} \xi_q^+ \tag{8}$$

Now let us tentatively assume that $\widehat{\Delta}\left(y^+, y_q^{+*}\right)$ measures the "neighborhood diversity" score difference between y^+ and the ground truth. Then Eq. 6 encourages any ranking predicted by W to have a neighborhood as diverse as the ground truth.

现在让我们暂时假设 $\widehat{\Delta}\left(y^+, y_q^{+*}\right)$ 测量 y^+ 和真实值之间的 "邻域多样性" 得分差异。那么公式 6 鼓励任何由 W 预测的排名具有与真实值一样多样的邻域。

To measure "neighborhood diversity" we introduce a new function:

为了测量 "邻域多样性"，我们引入一个新函数:

$$\operatorname{Div}\left(q, y^+, k'\right) = \frac{1}{k'} \sum_{i \in y^+, i=1}^{k'} \mathbf{1}\left(d_i \neq d_q\right) \tag{9}$$

which is the percentage of items not in d_q among the top- k' positions of y^+ . Based on Eq.9, the following diversity score function is devised:

这是 y^+ 的前 k' 个位置中不在 d_q 中的项目的百分比。基于公式 9，设计了以下多样性得分函数:

$$\operatorname{Div} S\left(q, y^+, k'\right) = \frac{1}{1 + e^{-\operatorname{Div}(q, y^+, k')/\eta}} \tag{10}$$

where η is, mathematically, a hyperparameter controlling the shape of this sigmoid-like function. Eq. 10 is needed because setting the ground truth y_q^{+*} to have the property in Eq. 7 may dangerously overconnect or overmerge datasets. The degree to which we can connect the data is data dependent. If the data resists merging, we need to set η to be a small value, so that even a tiny increase of Eq. 9 leads to a gigantic boost in Eq.10., resulting in a score very close to ground truth. At a higher level, η is a

hyperparameter that we should vary and one that depends on the intrinsic property of training data. Finally, a natural choice for $\widehat{\Delta}$ is the score difference:

其中 η 在数学上是一个超参数，用于控制这个类似于 sigmoid 的函数的形状。公式 10 是必要的，因为将真实值 y_q^{+*} 设置为具有公式 7 中的属性可能会危险地过度连接或过度合并数据集。我们可以连接数据的程度依赖于数据本身。如果数据抵制合并，我们需要将 η 设置为一个小值，以便即使公式 9 的微小增加也会导致公式 10 的巨大提升，从而得到一个非常接近真实值的得分。在更高的层面上，η 是一个我们应该变化的超参数，并且依赖于训练数据的内在特性。最后，$\widehat{\Delta}$ 的一个自然选择是得分差异:

$$\widehat{\Delta}\left(y^{+}, y_q^{+*}\right) = \operatorname{Div} S\left(q, y_q^{+*}, k'\right) - \operatorname{Div} S\left(q, y_q^{+}, k'\right). \tag{11}$$

Optimization When solving for the optimal W in Eq.8, we can not enumerate the entire $\mathcal{Y}$ and $\mathcal{Y}^+$ to list all the constraints. To optimize for UML, we use an efficient cutting-plane algorithm, which is slightly different from [15][11] due to the additional constraint of Eq.6. The general idea is that for each training example, maintain one set $\mathcal{S}_i$ of active constraints corresponding to Eq.2, and another set $\mathcal{S}_i^+$ of active constraints corresponding to Eq.6. The algorithm alternates between solving for W under current active constraints, and updating $\mathcal{S}_i$ and $\mathcal{S}_i^+$ by adding to them the most violated constraint $\widehat{y}_i$ and $\widehat{y}_i^+$ of each training example under current W . If a newly found $\widehat{y}_i\left(\widehat{y}_i^+\right)$ has a violation, which is the value of its slack variable indeed, greater than current violation $\xi_i\left(\xi_i^+\right)$ by some threshold ϵ , then it will be added, otherwise discarded. This iteration keeps going until no new constraint needs to be added. Gradient descent is used to solve for W , and the solution at each gradient step will be projected to its positive semidefinite (PSD) cone so that W is a feasible metric. Algorithm 1 is a high level illustration of the optimization procedure for UML.

优化在求解 Eq.8 中的最优 W 时，我们无法枚举整个 $\mathcal{Y}$ 和 $\mathcal{Y}^+$ 来列出所有约束。为了优化 UML，我们使用了一种高效的切平面算法，该算法由于 Eq.6 的附加约束与 [15][11] 略有不同。一般思路是，对于每个训练样本，维护一组与 Eq.2 对应的活动约束 $\mathcal{S}_i$ ，以及另一组与 Eq.6 对应的活动约束 $\mathcal{S}_i^+$ 。该算法在当前活动约束下交替求解 W ，并通过将当前 W 下每个训练样本的最违反约束 $\widehat{y}_i$ 和 $\widehat{y}_i^+$ 添加到 $\mathcal{S}_i$ 和 $\mathcal{S}_i^+$ 中来更新它们。如果新找到的 $\widehat{y}_i\left(\widehat{y}_i^+\right)$ 的违反程度，即其松弛变量的值，超过当前违反程度 $\xi_i\left(\xi_i^+\right)$ 一定的阈值 ϵ ，则将其添加，否则丢弃。该迭代过程持续进行，直到不再需要添加新约束。梯度下降法用于求解 W ，每个梯度步骤的解将被投影到其正半定 (PSD) 锥中，以确保 W 是一个可行的度量。算法 1 是 UML 优化过程的高层次示意图。

Notice that in order to further speed up the optimization process, two modification are made in our implementation. First, as in [15][11] the problem is reformulated so that only one or two global slack variables are maintained. Second, as in [13], we use the alternating direction method of multipliers [2] to reduce the number of times eigen-decomposition is performed, which is done when projecting the solution to its PSD cone.

请注意，为了进一步加速优化过程，我们的实现中进行了两项修改。首先，如 [15][11] 中所述，问题被重新表述，以便仅维护一个或两个全局松弛变量。其次，如 [13] 中所述，我们使用交替方向乘子法 [2] 来减少特征分解的次数，这在将解投影到其 PSD 锥时进行。

Algorithm 1 Optimization Algorithm

Input: training data $\mathcal{X}$, positive/negative set $\mathcal{X}_i^+/\mathcal{X}_i^-$, ranking y_i^* and y_i^{+*} , hyperparamters: $C_1 > 0, C_2 > 0$ and $\eta > 0$, stop threshold $\epsilon > 0$

Output: $W \succcurlyeq 0$ and slack variables ξ_i and ξ_i^+

for all $i = 1, \ldots, n, \mathcal{S}_i \leftarrow \varnothing, \mathcal{S}_i^+ \leftarrow \varnothing, \xi_i \leftarrow 0, \xi_i^+ \leftarrow 0$

repeat

for $i = 1, \ldots, n$ do

$\widehat{y}_i \leftarrow \operatorname{argmax}_{y \in \mathcal{Y}} \langle W, \psi_{po}(i, y) \rangle_F + \Delta(y, y_i^*)$

$\widehat{y}_i^+ \leftarrow \operatorname{argmax}_{y^+ \in \mathcal{Y}^+} \langle W, \psi_{po}(i, y^+) \rangle_F + \widehat{\Delta}\left(y^+, y_i^{+*}\right)$

if the slack of $\widehat{y}_i$ is greater than $\xi_i + \epsilon$ then

$\mathcal{S}_i \leftarrow \mathcal{S}_i \cup \{\widehat{y}_i\}$

end if

if the slack of $\widehat{y}_i^+$ is greater than $\xi_i^+ + \epsilon$ then

$\mathcal{S}_i^+ \leftarrow \mathcal{S}_i^+ \cup \left\{\widehat{y}_i^+\right\}$

end if

Solve for Eq. 8

subject to $\mathcal{S}_i$ and $\mathcal{S}_i^+$ for all $i = 1, \ldots, n$
end for
until no $\mathcal{S}_i$ and $\mathcal{S}_i^+$ has changed during iteration

---

5. Experiments

5. 实验

In this section, we evaluate our approach on the image classification task.

在本节中，我们评估我们的方法在图像分类任务上的表现。

5.1. Data

5.1. 数据

We used four standard datasets: Caltech101 [7], PASCAL VOC [4], LabelMe [20] and SUN09 [3]. The following five common categories were selected: bird, car, chair, dog and person.

我们使用了四个标准数据集:Caltech101 [7]、PASCAL VOC [4]、LabelMe [20] 和 SUN09 [3]。选择了以下五个常见类别: 鸟、汽车、椅子、狗和人。

Training set The training set contains images from the four datasets. For each category in each dataset, we randomly selected up to 100 images, so that we could vary the number of training examples per category per dataset. In cross-dataset classification experiments one dataset will be held out as unseen dataset. Therefore, we remove its images from the pool of training images to form a subset (denoted as $\mathbf{Tr}^{\text{unseen}}$) and the rest of the images form another subset called Tr^{seen} . In practice, only Tr^{seen} was used

训练集训练集包含来自四个数据集的图像。对于每个数据集中的每个类别，我们随机选择最多 100 张图像，以便能够在每个数据集的每个类别中变化训练样本的数量。在跨数据集分类实验中，将保留一个数据集作为未见数据集。因此，我们从训练图像池中移除其图像，以形成一个子集 (记作 $\mathbf{Tr}^{\text{unseen}}$)，其余图像形成另一个子集，称为 Tr^{seen} 。实际上，仅使用了 Tr^{seen} 。

Validation set This validation set contains images from the four datasets. The 20 images per category per dataset are randomly selected. Similar to the training set, in cross-dataset experiments one dataset will be held out, so we have two subsets as $\text{Va}^{\text{unseen}}$ and Va^{seen} .

验证集该验证集包含来自四个数据集的图像。每个类别每个数据集随机选择 20 张图像。与训练集类似，在跨数据集实验中，将保留一个数据集，因此我们有两个子集，分别为 $\text{Va}^{\text{unseen}}$ 和 Va^{seen} 。

Web validation set In order to carry out our novel validation method, we constructed another validation set with web images. We issued the class labels as keyword to Google and downloaded the top 50 returned images. After removing duplicates with regard to the training set, 20 images were randomly selected for each class to form the set. We will refer to it as Va^{web} .

网络验证集为了实施我们新颖的验证方法，我们构建了另一个包含网络图像的验证集。我们将类别标签作为关键词输入 Google，并下载返回的前 50 张图像。在与训练集进行重复性检查后，为每个类别随机选择 20 张图像以形成该集合。我们将其称为 Va^{web} 。

Test set There are 20 images per class per dataset. Similar to training set, the test set was divided into $\mathbf{Te}^{\text{unseen}}$ and Te^{seen} in cross-dataset evaluation.

测试集每个类别每个数据集有 20 张图像。与训练集类似，测试集在跨数据集评估中被划分为 $\mathbf{Te}^{\text{unseen}}$ 和 Te^{seen} 。

We generated 3 different copies of the above sets, so that all of our following experiments are carried out 3 times on different data and the mean results are reported. For each image, grayscale SIFT descriptors [14] were extracted at interest points detected with the Hessian-Affine detector [16]. Then we quantized these descriptors to bag-of-words representation using a vocabulary of size 500 at 3 spatial pyramid levels and the feature representation is of 10,500 dimensions. Eventually, PCA was applied to reduce the dimensionality to 800 .

我们生成了上述数据集的 3 个不同副本，以便我们后续的所有实验在不同的数据上进行 3 次，并报告平均结果。对于每个图像，使用 Hessian-Affine 检测器 [16] 检测到的兴趣点提取了灰度 SIFT 描述符 [14]。然后，我们使用大小为 500 的词汇在 3 个空间金字塔层次上对这些描述符进行量化，特征表示的维度为 10,500。最终，应用 PCA 将维度降低到 800。

We started with an experiment to validate the existence of bias in our datasets. Then we applied our method to cross-dataset classification task. Finally, we compare our approach with other metric learning methods. A simple k - nearest neighbor classifier is used in all classification experiments. Details and results are shown in the following sections.

我们首先进行了一项实验，以验证我们的数据集中存在偏差。然后，我们将我们的方法应用于跨数据集分类任务。最后，我们将我们的方法与其他度量学习方法进行了比较。在所有分类实验中使用了简单的 k - 最近邻分类器。详细信息和结果将在以下章节中展示。

5.2. Existence of bias

5.2. 偏差的存在

In this first experiment we show that our four datasets are strongly biased, in the sense that a model trained on one dataset is ineffective (due to bias) on the other datasets. We do this by using SML to learn a metric on each dataset individually and then test it on each dataset individually. For each metric, the classification accuracy on each dataset is reported. We used 20 images per category per dataset to form the training set, as well as the validation and test sets. The hyperparameter C in Eq. 1 was selected from the following values $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$.

在第一次实验中，我们展示了我们的四个数据集存在明显的偏差，意味着在一个数据集上训练的模型在其他数据集上效果不佳 (由于偏差)。我们通过使用 SML 在每个数据集上单独学习一个度量，然后在每个数据集上单独测试来实现这一点。对于每个度量，报告每个数据集上的分类准确率。我们使用每个类别每个数据集的 20 张图像来形成训练集，以及验证集和测试集。方程 1 中的超参数 C 从以下值中选择 $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ 。

Train	Test			
	Cal	Pas	SUN	Lab
Cal	0.87	0.33	0.24	0.39
Pas	0.31	0.40	0.32	0.30
SUN	0.11	0.23	0.37	0.22
Lab	0.24	0.25	0.18	0.47

训练	测试			
	校准	通过	SUN	实验室
校准	0.87	0.33	0.24	0.39
通过	0.31	0.40	0.32	0.30
SUN	0.11	0.23	0.37	0.22
实验室	0.24	0.25	0.18	0.47

Table 1. Classification accuracy on all datasets. Metrics are learned on each dataset individually. The left-most column specifies the training dataset where the metric is learned, while the uppermost row specifies the test dataset. Cal, Pas, SUN and Lab stand for Caltech101, PASCAL VOC, SUN09 and LabelMe respectively.

表 1. 所有数据集上的分类准确率。度量是在每个数据集上单独学习的。最左侧的列指定了学习度量的训练数据集，而最上面的行指定了测试数据集。Cal、Pas、SUN 和 Lab 分别代表 Caltech101、PASCAL VOC、SUN09 和 LabelMe。

Table 1 shows clearly that our datasets are highly biased. With the same test example, the performance of models trained on different datasets varies widely (read Table 1 vertically). For example, when tested on Caltech101 (see column 1), the best classification performance is achieved by the model trained on the same dataset, while the lowest accuracy comes from the model trained on SUN09 images.

表 1 清楚地显示了我们的数据集存在高度偏差。在相同的测试示例下，训练于不同数据集的模型性能差异很大 (请垂直阅读表 1)。例如，在 Caltech101 上测试时 (见第 1 列)，最佳分类性能来自于在相同数据集上训练的模型，而最低准确率则来自于在 SUN09 图像上训练的模型。

5.3. Cross-dataset classification

5.3. 跨数据集分类

In this experiment, our goal is to measure the true generalizability of the model learned by UML with cross-dataset performance. We should point out that cross-dataset classification does not completely measure a model's true generalizability, but it is a reasonable indicator.

在本实验中，我们的目标是通过跨数据集性能来测量 UML 学习模型的真实泛化能力。我们应该指出，跨数据集分类并不能完全衡量模型的真实泛化能力，但它是一个合理的指标。

Since the performance of UML is subject to model training as well as validation, our experiment is composed of two stages, where we evaluate them separately. Finally, we will do full evaluation and compare UML to other metric learning methods.

由于 UML 的性能受模型训练和验证的影响，我们的实验分为两个阶段，分别进行评估。最后，我们将进行全面评估，并将 UML 与其他度量学习方法进行比较。

5.3.1 Training stage

5.3.1 训练阶段

In this experiment, we want to test the effectiveness of the UML learning framework, so the question for this stage is whether UML can produce models with better generalizability. Recall that by varying hyperparameters we will learn a set of metrics with different levels of "neighborhood label coherence" and "neighborhood diversity". These metrics will be validated by the validation procedure. However, if the validation is carried out improperly (e.g., using a validation set that is not related to the target task) then a bad metric can be selected. For example, in our case, it would be bad to use $\mathrm{Va}^{\mathrm{seen}}$, since the goal is to generalize to every possible unseen datasets/domains, not only the three seen datasets. The corresponding consequence is that it is impossible to tell the quality of models produced by UML training step. Clearly, in this stage, we need a validation set that is closely related to the target task, which is, in the cross-dataset setting, generalizing to the unseen dataset. The choice is obvious: using $\mathrm{Va}^{\mathrm{unseen}}$ as the validation set at this stage. In this way, the effect of the validation procedure on the final performance is minimized, thus resulting in a better evaluation of UML's learning step.

在本实验中，我们希望测试 UML 学习框架的有效性，因此本阶段的问题是 UML 是否能够生成具有更好泛化能力的模型。回想一下，通过改变超参数，我们将学习一组具有不同“邻域标签一致性”和“邻域多样性”水平的指标。这些指标将通过验证程序进行验证。然而，如果验证不当 (例如，使用与目标任务无关的验证集)，则可能选择到不好的指标。例如，在我们的案例中，使用 $\mathrm{Va}^{\mathrm{seen}}$ 是不合适的，因为目标是对每个可能的未见数据集/领域进行泛化，而不仅仅是三个已见数据集。相应的后果是，无法判断 UML 训练步骤所产生模型的质量。显然，在这个阶段，我们需要一个与目标任务密切相关的验证集，即在跨数据集设置中，对未见数据集的泛化。选择显而易见: 在这个阶段使用 $\mathrm{Va}^{\mathrm{unseen}}$ 作为验证集。通过这种方式，验证程序对最终性能的影响被最小化，从而更好地评估 UML 的学习步骤。

In detail, when a dataset was marked as unseen, the corresponding $\mathrm{Tr}^{\mathrm{seen}}$ and $\mathbf{Te}^{\mathrm{unseen}}$ denote the training set and test set respectively. There were 20 images per class per dataset in $\mathrm{Va}^{\mathrm{unseen}}$ and we let the number of training examples per class per dataset in $\mathrm{Tr}^{\mathrm{seen}}$ to be the following values $\{15, 20, 30, 50, 70, 100\}$. The validation set, as explained above, was $\mathrm{Va}^{\mathrm{unseen}}$, which was of the same size as $\mathbf{Te}^{\mathrm{unseen}}$. The k in Eq. 5 and the k' in Eq. 11 was set to be the same as current number of training examples. The neighborhood size of the k -nearest neighbor classifier was determined via validation. We let each dataset be unseen once, and report in Fig. 2 the test accuracy on the corresponding $\mathbf{Te}^{\mathrm{unseen}}$, as well as the average performance of these individual experiments. From the top figure in Fig. 2 we can tell that the models produced by UML have better cross-dataset generalizability than SML. This suggest that utilizing domain information (such as the source of the example dataset), in learning can be helpful to extract more general knowledge. The bottom figures report the results of individual experiments. As shown, UML still gives better results most times. One observation we found is that UML has fewer advantages over baseline when fewer (e.g., 15) training examples are used, but benefits more from more training examples. This is because UML can not enforce "neighborhood diversity" due to the lack of good pairs to pull together, and with more examples at hand, better pairs can be found and this in turn give rise to "better" neighborhoods can be constructed. There are certain points at which both methods do not benefit from more training examples. This is due to the fact that newly added examples can bring in more bias, which will then further bias the derived learning algorithm.

具体来说，当一个数据集被标记为未见时，相应的 $\mathrm{Tr}^{\mathrm{seen}}$ 和 $\mathbf{Te}^{\mathrm{unseen}}$ 分别表示训练集和测试集。在 $\mathrm{Va}^{\mathrm{unseen}}$ 中每个类别每个数据集有 20 张图像，我们将每个类别每个数据集的训练样本数量设为以下值 $\{15, 20, 30, 50, 70, 100\}$ 。如上所述，验证集为 $\mathrm{Va}^{\mathrm{unseen}}$ ，其大小与 $\mathbf{Te}^{\mathrm{unseen}}$ 相同。公式 5 中的 k 和公式 11 中的 k' 设置为当前训练样本数量相同。通过验证确定 k -最近邻分类器的邻域大小。我们让每个数据集未见一次，并在图 2 中报告相应 $\mathbf{Te}^{\mathrm{unseen}}$ 的测试准确率，以及这些单独实验的平均性能。从图 2 的顶部图形中可以看出，UML 产生的模型在跨数据集的泛化能力上优于 SML。这表明，在学习中利用领域信息 (例如示例数据集的来源) 有助于提取更一般的知识。底部图形报告了单独实验的结果。如所示，UML 在大多数情况下仍然提供更好的结果。我们发现的一个观察是，当使用较少 (例如 15) 训练样本时，UML 相对于基线的优势较小，但在使用更多训练样本时受益更多。这是因为 UML 由于缺乏良好的配对而无法强制执行 “邻域多样性”，而手头有更多示例时，可以找到更好的配对，从而构建出 “更好” 的邻域。在某些情况下，两种方法都不会从更多的训练样本中受益。这是因为新添加的示例可能带来更多的偏差，从而进一步偏向所推导的学习算法。

5.3.2 Validation stage

5.3.2 验证阶段

The goal of the second stage is to evaluate how different validation sets affect the final performance. Two validation sets were used, $\mathrm{Va}^{\mathrm{web}}$ and the corresponding $\mathrm{Va}^{\mathrm{seen}}$ in individual experiments. The set of metrics produced by UML in the training stage was the input, thus with the same input, our expectation was that our validation procedure with $\mathrm{Va}^{\mathrm{web}}$ would consistently outperform the other. This is proved by results reported in Fig.3, in which metrics selected by $\mathrm{Va}^{\mathrm{web}}$ not only outperform metrics selected by $\mathrm{Va}^{\mathrm{seen}}$, but almost matches the ones selected by $\mathrm{Va}^{\mathrm{unseen}}$.

第二阶段的目标是评估不同验证集对最终性能的影响。使用了两个验证集，$\mathrm{Va}^{\mathrm{web}}$ 和在各个实验中相应的 $\mathrm{Va}^{\mathrm{seen}}$ 。在训练阶段由 UML 产生的指标集作为输入，因此在相同输入的情况下，我们期望使用 $\mathrm{Va}^{\mathrm{web}}$ 的验证程序能够始终优于其他方法。这一点在图 3 中的结果得到了证明，其中由 $\mathrm{Va}^{\mathrm{web}}$ 选择的指标不仅优于由 $\mathrm{Va}^{\mathrm{seen}}$ 选择的指标，几乎与由 $\mathrm{Va}^{\mathrm{unseen}}$ 选择的指标相匹配。

An additional question we ask is whether web images can help conventional metric learning method to generalize. To answer it, we applied $\mathrm{Va}^{\mathrm{seen}}$ and $\mathrm{Va}^{\mathrm{web}}$ to metrics produced by SML, and found that metrics selected by $\mathrm{Va}^{\mathrm{web}}$ outperform those of $\mathrm{Va}^{\mathrm{seen}}$. Due to the lack of space, we leave the corresponding figure in the supplementary material.

我们提出的另一个问题是，网络图像是否可以帮助传统的度量学习方法进行泛化。为了回答这个问题，我们将 $\mathrm{Va}^{\mathrm{seen}}$ 和 $\mathrm{Va}^{\mathrm{web}}$ 应用于 SML 产生的指标，并发现由 $\mathrm{Va}^{\mathrm{web}}$ 选择的指标优于 $\mathrm{Va}^{\mathrm{seen}}$ 的指标。由于空间有限，我们将相应的图表留在补充材料中。

5.3.3 Full evaluation

5.3.3 完整评估

In this experiment, we compared our approach, UML, to the following baseline methods: Structural Metric Learning (SML) [15], Large Margin Nearest Neighbor (LMNN) [24], Maximally Collapsing Metric Learning (MCML) [9] and Large Margin Multi-Task Metric Learning (mtLMNN) [17]. In mtLMNN, a task was to classify on one of seen training sets, and the learned shared metric, which encodes the knowledge shared among different tasks, was applied to unseen test set. We vary hyperparameters for all baseline methods and validate the produced metrics on $\mathrm{Va}^{\mathrm{seen}}$. As Fig. 4 demonstrates our approach significantly outperforms all the baseline methods.

在这个实验中，我们将我们的方法 UML 与以下基线方法进行了比较: 结构度量学习 (SML) [15]、大边距最近邻 (LMNN) [24]、最大崩溃度量学习 (MCML) [9] 和大边距多任务度量学习 (mtLMNN) [17]。在 mtLMNN 中，一个任务是在已见训练集上进行分类，学习到的共享度量编码了不同任务之间共享的知识，并应用于未见测试集。我们为所有基线方法调整超参数，并在 $\mathrm{Va}^{\mathrm{seen}}$ 上验证产生的指标。如图 4 所示，我们的方法显著优于所有基线方法。

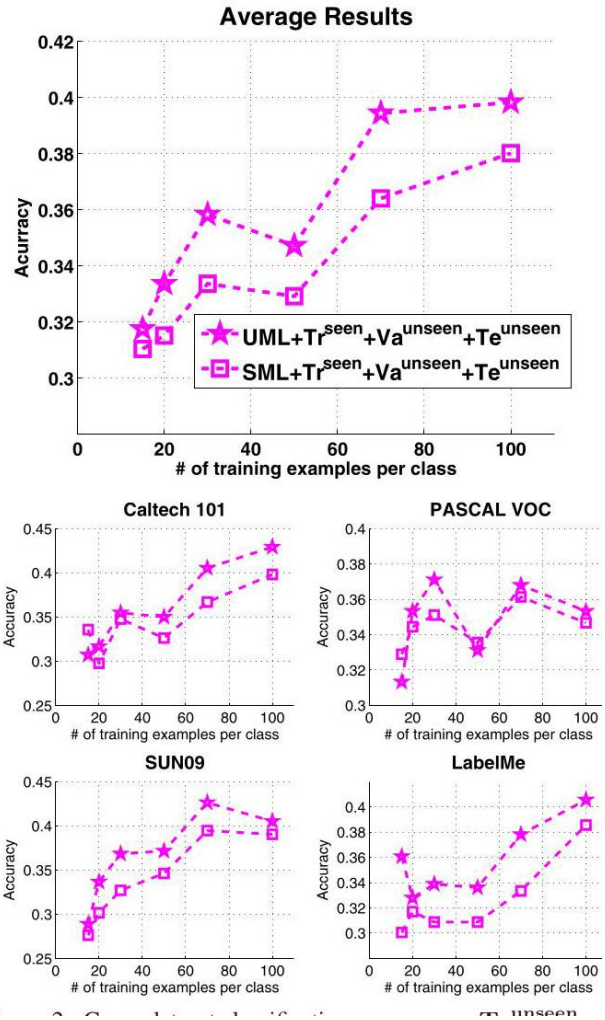

Figure 2. Cross-dataset classification accuracy on Te $^{\text{unseen}}$ with validation on Va $^{\text{unseen}}$. Top: the figure reports the average accuracy over four individual experiments in the bottom. Bottom: individual classification accuracy on hold-out dataset.

图 2. 在 Te $^{\text{unseen}}$ 上的跨数据集分类准确率，验证集为 Va $^{\text{unseen}}$ 。上部分: 该图报告了底部四个独立实验的平均准确率。下部分: 在保留数据集上的个体分类准确率。

6. Discussion

6. 讨论

The goal of this paper is to learn a less biased distance metric that generalizes better to unseen datasets. First, we introduced the notion of "neighborhood diversity" to better connect examples from different datasets and extract general knowledge. We then proposed to use web images as a validation set to select metrics with better generalizability. We have shown that our approach is able to produce superior performance in cross-dataset image classification experiments on four popular datasets.

本文的目标是学习一种偏差较小的距离度量，以更好地推广到未见过的数据集。首先，我们引入了"邻域多样性"的概念，以更好地连接来自不同数据集的示例并提取一般知识。然后，我们提出使用网络图像作为验证集，以选择具有更好泛化能力的度量。我们已经证明，我们的方法能够在四个流行数据集上产生优越的跨数据集图像分类实验性能。

The contributions of this paper are two-fold. We demonstrated that by tuning the distribution of data from different domains, more generalizable models can be produced. We also showed the advantage of using weakly-labeled web images as validation set to select model with better generalization ability. Web images are known to be easy and cheap to obtain. Our work uncovers another nice property: less biased.

本文的贡献有两个方面。我们展示了通过调整来自不同领域的数据分布，可以生成更具泛化能力的模型。我们还展示了使用弱标记的网络图像作为验证集以选择具有更好泛化能力的模型的优势。网络图像被认为容易且廉价获取。我们的工作揭示了另一个良好的特性: 偏差较小。

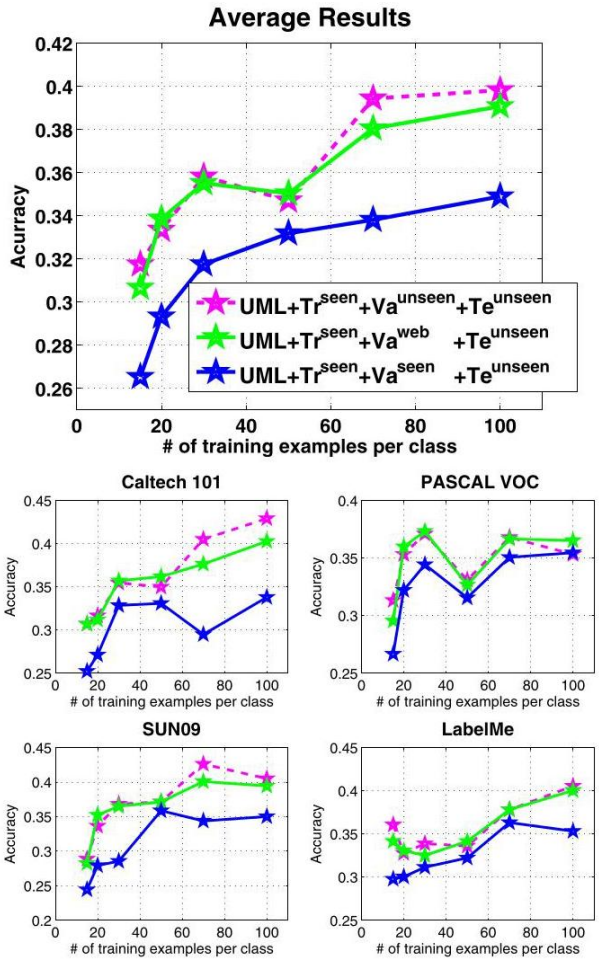

Figure 3. Cross-dataset classification accuracy on unseen test sets. Different validation methods are compared. Top: the figure reports the average accuracy over four individual experiments in the bottom. Bottom: individual classification accuracy on hold-out dataset. The green curve is our validation method using Va^{web} and the blue one is conventional validation using Va^{seen} . The magenta dash line is validated on $\text{Va}^{\text{unseen}}$ (the same as in Fig.2).

图 3. 在未见测试集上的跨数据集分类准确率。比较了不同的验证方法。上部分: 该图报告了底部四个独立实验的平均准确率。下部分: 在保留数据集上的个体分类准确率。绿色曲线是我们使用 Va^{web} 的验证方法，蓝色曲线是使用 Va^{seen} 的传统验证。品红色虚线是在 $\text{Va}^{\text{unseen}}$ 上验证的 (与图 2 中相同)。

Acknowledgements

致谢

We are grateful to Aditya Khosla for sharing data. Thanks to Alessandro Bergamo and Yuting Sun for proofreading drafts. The authors were partly supported by AFOSR Award FA9550-11-1-0166 and the Neukom Institute for Computational Science.

我们感谢 Aditya Khosla 分享数据。感谢 Alessandro Bergamo 和 Yuting Sun 校对草稿。作者部分得到了 AFOSR 奖项 FA9550-11-1-0166 和 Neukom 计算科学研究所的支持。

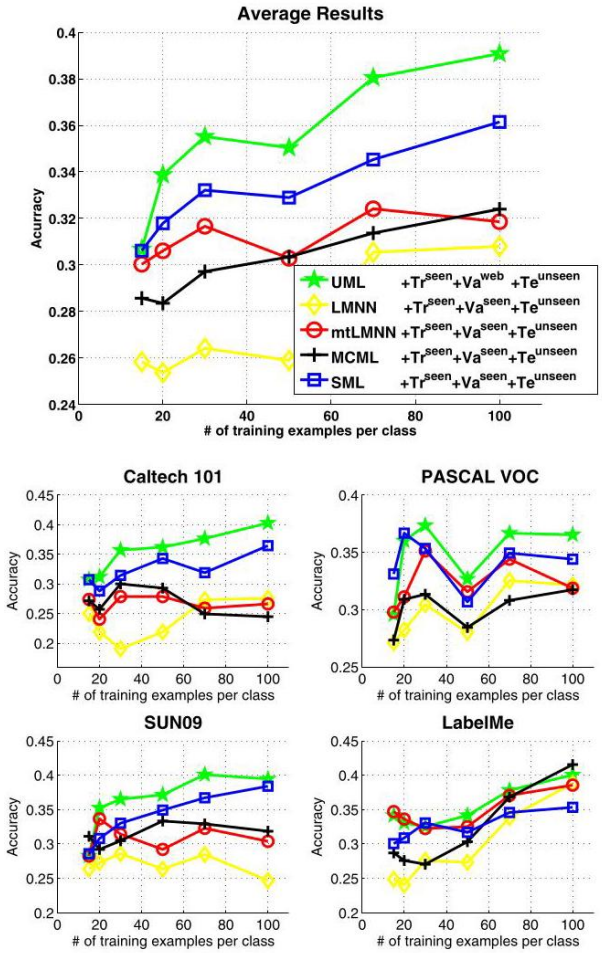

Figure 4. Cross-dataset classification accuracy on unseen test sets. UML (green curve) is compared with baselines. Top: the average accuracy over four individual experiments in the bottom. Bottom: individual classification accuracy on hold-out dataset.

图 4. 在未见测试集上的跨数据集分类准确率。UML(绿色曲线) 与基线进行比较。顶部: 底部四个独立实验的平均准确率。底部: 在保留数据集上的个体分类准确率。

References

参考文献

[1] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou, and J. S. Marron. Adjustment of systematic microarray data biases. Bioinformatics, 20(1):105-114, 2004.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn., 3(1):1-122, Jan. 2011.

[3] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In CVPR, 2010.

[4] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (voc) challenge. IJCV, 88(2):303-338, June 2010.

[5] T. Evgeniou and M. Pontil. Regularized multi-task learning. In ACM SIGKDD, pages 109-117, 2004.

[6] C. Fang and L. Torresani. Measuring image distances via embedding in a semantic manifold. In ECCV, 2012.

[7] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Underst.,

106(1):59–70, Apr. 2007.

[8] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In ICCV, 2005.

[9] A. Globerson and S. Roweis. Metric learning by collapsing classes. NIPS, 2006.

[10] L. jia Li, G. Wang, and L. Fei-fei. Optimol: automatic online picture collection via incremental model learning. In CVPR, 2007.

[11] T. Joachims. A support vector method for multivariate performance measures. In ICML, 2005.

[12] A. Khosla, T. Zhou, T. Malisiewicz, A. Efros, and A. Tor-ralba. Undoing the damage of dataset bias. In ECCV, 2012.

[13] D. Lim, B. Mcfee, and G. R. Lanckriet. Robust structural metric learning. In ICML, 2013.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91-110, Nov. 2004.

[15] B. Mcfee and G. Lanckriet. Metric learning to rank. In ICML, 2010.

[16] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. IJCV, 60(1):63-86, 2004.

[17] S. Parameswaran and K. Weinberger. Large margin multitask metric learning. In NIPS. 2010.

[18] W. Ping, Y. Xu, J. Wang, and X.-S. Hua. Famer: Making multi-instance learning better and faster. In SDM, 2011.

[19] M. Rastegari, C. Fang, and L. Torresani. Scalable object-class retrieval with approximate and top-k ranking. In ICCV, 2011.

[20] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. IJCV, 77(1-3):157-173, May 2008.

[21] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In ECCV, 2010.

[22] T. Tommasi, N. Quadrianto, B. Caputo, and C. H. Lampert. Beyond dataset bias: Multi-task unaligned shared knowledge transfer. In ACCV, 2012.

[23] A. Torralba and A. Efros. Unbiased look at dataset bias. In CVPR, 2011.

[24] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res., 10:207-244, June 2009.

[25] Y. Xu, W. Ping, and A. Campbell. Multi-instance metric learning. In ICDM, 2011.

[26] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In CVPR, 2009.

[27] Y. Yue and T. Finley. A support vector method for optimizing average precision. In SIGIR, 2007.