

Label-Embedding for Image Classification

一. 写作背景及目的

1. 背景

在图像分类问题中，由于不可能对所有类别的物体进行标记，所以零样本学习具有很高的实用价值。零样本学习的解决方案中，有一种方案被称为基于属性的学习，基于属性的学习引入了一个称为属性层的中间层，通常这些属性由机器检测并可以被人所理解。根据每个类别对应的属性是否存在，可以将每个类表示为类-属性关联向量。

2. 目的

提出一种基于标签嵌入框架，同时可以方便利用辅助信息的零样本图像分类方案。

二. 核心问题

为了对新图像进行分类，基于属性的预测方法首先需要为每一个属性学习一个分类器，然后再使用学习到的分类器预测其属性，最后再将每个属性的分数合并成对应类别的分数，以此来进行预测类别。此方法被称为直接属性预测(DAP)。DAP主要存在以下几点问题：

1. 由于属性分类器独立于最终任务而学习的，所以可能在预测属性时性能是最佳的，但是在预测类别时却没法达到最佳性能。
2. DAP很难在增量学习情景下利用新标注的样本。
3. 属性信息获得成本很高，而且人类标注的信息不一定很可靠。
4. DAP很难利用一些新增的辅助信息(如文本描述信息)。

三. 现有状况及相关研究

1. 属性

当前流行的基于属性的识别方案是直接预测模型。此方法首先预测图像中是否存在某些属性，再将这些属性的预测概率组合成类别预测概率。DAP的一个重要限制是假设属性之间相互独立，但是这个假设往往是错误的。

2. 零样本学习

零样本学习要求方法具有将知识从参与训练类别中转移到没有参与训练的类别中。在进行零样本学习时，先验信息的选择和识别模型的选择是比较重要的两个选择。目前先验信息的可能来源主要包括属

性, 类间相似性, 文本特征和类共现统计等。对于模型来说, 目前也有几种选择, 一种是前面提到的基于属性独立假设的**DAP**模型, 还有就是采用度量图像与类嵌入间距离进行分类的模型。

3. 标签嵌入

在计算机视觉中, 有着大量的相关研究致力于如何更好的表示一幅图像, 即输入嵌入。常见的基于内核的方法, 以及基于显性嵌入, 降维和压缩等的方法。而标记嵌入方面相关的研究却比较少。只要正确选择嵌入函数 ϕ , 即在嵌入空间中, 相似类的标记嵌入的欧几里得度量是接近的, 则标记嵌入就能成为实现类间参数共享的有效方法。其主要可以用于多分类, 零样本学习等方面。作者在此主要关注输出嵌入, 输出嵌入可以是固定且与数据无关的, 也可以是从数据中学习出来的, 或者从辅助信息中计算出来的。

3.1 数据无关嵌入

内核依赖评估是一种 ϕ 独立于数据的方法, 其中 ϕ 通过 Y 空间中的内核隐式定义, 而 Hsu 等人的压缩感知方法也是一种 ϕ 独立于数据的方法, 其中 ϕ 对应于随机投影。

3.2 学习嵌入

此方法主要是联合学习 θ 和 ϕ , 将输入和输出嵌入到一个共同的中间空间 Z 。最典型的例子是典型相关分析(CCA), 它最大限度地提高了输入和输出之间的相关性。还有一些其他策略最大限度地提高了分类精度, 包括 AMIT 等的核规范正则化学习和 Weston 等人的 WSABIE 算法。

3.3 从侧面信息导出的嵌入

当可用于训练的数据很少时, 侧面信息可以弥补数据缺乏的问题。Frome 等人使用了文本语料库去进行类嵌入, 然后使用排序目标函数学习输入与输出嵌入之间的映射。

作者在本文中主要采用了从零样本识别的侧面信息导出的嵌入, 同时也考虑了数据无关的嵌入和学习嵌入。

四. 作者所做的工作

1. 基本框架

作者定义了一个双线相容函数 F 将视觉信息和辅助信息关联起来:

$$F(x, y; W) = \theta(x)' W \varphi(y)$$

定义这个函数的目的是对于见过或没见过的类别, 衡量图像特征 $\theta(x)$ 和语义表征等辅助信息 $\varphi(y)$ 之间的相容性 (compatibility)。 W 是所要学习的视觉-语义映射矩阵, 大小为 $D \times E$ 维。如果 D 和 E 比较大, 作者考虑采用一个低秩分解的方式将 W 表示成了 U 与 V 两个矩阵的形式, 来降低参数量。这时 F 函数就变成了:

$$F(x, y; U, V) = (U \theta(x))' (V \phi(y))$$

2. 带有属性的标签嵌入

作者假设了在每个属性 a_i 和每个类 y 之间存在一个关联度量 $\rho_{y,i}$,则可以将 y 通过如下的公式嵌入到一个 E 维的空间中.

$$\phi^A(y) = [\rho_{y,1}, \dots, \rho_{y,E}]$$

作者在对输出向量规范化的过程中,比较了(1)连续嵌入, (2)编码使用 $\{0, 1\}$ 的二进制嵌入, (3)使用 $\{-1, 1\}$ 编码的二进制嵌入等方式. 同时作者还探讨了两种归一化策略: (1)均值中心(即计算所有参与训练类别的平均值并减去它)和(2)L2-归一化.

在属性比较多余的情况下,更好的方式是对他们进行去相关操作, 作者在实验中研究了属性去相关的影响。

3. 属性之外的标签嵌入

作者提出了一个包含所有以前的关于标签嵌入方法的总体框架, 并对它们在图像分类任务的经验性能进行了比较. 标签嵌入方法可以根据两个标准进行分类: (1)以任务为中心或使用其他来源的辅助信息; (2)固定的或依赖于任务数据的嵌入。

3.1 标签嵌入中的辅助信息

区分不同的标签嵌入方法的第一个标准是, 该方法是只使用目前任务的训练数据, 即示例(图像)及其类标签, 还是使用其他信息来源. 在后一种方案中, 不同的辅助信息可以使输出达到不一样的效果, 具体的辅助信息可以是: (1)属性, (2)类分类法,(3)文本语料库。

- 类层次结构显式地使用专家知识将图像类分组为层次结构, 例如鸟类数据集中来自鸟类学的知识。类的层次结构需要 Y 中的排序操作 $<$: $Z < y$ 意味着 z 是树层次结构中 y 的祖先。作者定义了如果 $z < y$ 或者 $z = y$, 则 $\xi_{y,z} = 1$, 否则为0. 则通过这种方式, 可以将类的层次嵌入 $\phi^H(y)$ 定义为

$$\phi^H(y) = [\xi_{y,1}, \dots, \xi_{y,C}]$$

作者将这种嵌入称为层次标签嵌入(HLE).

- 语料中类别的共现信息可以用来推断类之间的关系, 从而导出类的嵌入。类别的共现信息可以使用公共资源(如Wikipedia 2)自动提取。类名的共现作者使用了目前最先进的word2vec方法来导出标签嵌入, 并将其称之为WLE.

作者在第五节中, 比较了属性、类层次结构和文本信息(即ALE、HLE和WLE)作为零样本识别的辅助信息来源所能达到的效果。

3.2 标签嵌入中的数据相关性

区分不同的标签嵌入方法第二个标准是预测时使用的标签嵌入是否适合于训练时的训练数据。这里的数据相关性指的是训练数据中的相关性，而所有其他可能的信息来源则不管。这时可以分为三种方法：(一)固定和独立于数据的标签嵌入；(二)数据相关，仅从训练数据中学习；(三)数据相关，同时从训练数据和附加信息中共同学习。其中固定的和数据无关的方法对应于原始类标签到低维空间的固定映射。

在实验中，作者探索了三种嵌入：(1)与恒等映射相对应的一对多嵌入(OVR)；(2)高斯标签嵌入(GLE)，使用了高斯随机投影矩阵；(3)Hadamard标签嵌入，类似地，使用Hadamard矩阵构造随机投影矩阵。这三种标签嵌入方法中没有一种使用训练数据(或任何辅助信息)来构建标签嵌入。而数据相关的标签嵌入使用训练数据构建标签嵌入。

五. 实验验证

作者在两个公开数据集上评估了ALE的性能，分别是带有属性的动物数据集(AWA)和CUB-200-2011数据集(CUB)。AWA包含50个动物类的大约30,000幅图像。CUB包含大约11,800张200种鸟类的图片。

1. 输入嵌入

作者在不同尺度上通过规则网格提取了128维的SIFT特征和96维的色彩特征。再通过PCA将他们分别降到64维。再将这些特征转换为Fisher向量(FV)，作为图像级的表示。

2. 输出嵌入

实验中，作者考虑了三种可以导出嵌入的辅助信息：属性、类分类法和文本语料库。在考虑属性时，作者使用了二进制属性或连续属性。

- 基于属性的标签嵌入(ALE)

在AWA中，每个类别有85个属性标注，每个标注都有10个学生去完成，连续的类属性关联可以通过平均每个学生的选票来获得的，然后通过阈值化以获得二进制属性。在CUB中，可以从野外鸟类指南中获得312个属性，然后对于每一幅图像，根据是否存在某一属性来进行标注。

- 基于层次结构信息的标签嵌入(HLE)

作者使用了Wordnet层级信息来作为先验信息计算输出嵌入，作者使用Wordnet从AWA的50个类别中构建了一个含有150个节点的层级结构，然后对于一个给定的类，根据其祖先中是否存在相应的节点，将输出设置为0或者1，同时输出嵌入也做了L2规格化操作。在CUB数据集上作者也进行了同样的操作。

- Word2Vec标签嵌入(WLE)

作者在标记了150单词和短语的英语维基百科上训练了skip-gram模型，这些单词或者短语包含了之前提到的数据集里的视觉对象。

- 一对多嵌入(OVR)

OVR嵌入的维数和类别数相同，都为C。此时相当于为每个类别都训练一个分类器。

- 高斯标签嵌入(GLE)

GLE的类别嵌入是从标准正太分布中提取的，类似与压缩感知中的随机投影。

- Hadamard标签嵌入

在作者的实验中，Hadamard嵌入的结果比GLE明显差。因此，后面只分析GLE的结果。

- WSABIE

当有足够数量的标注训练数据时，可以从训练数据中学习嵌入。作者考虑了一种基于数据驱动的标签嵌入方法WSABIE，WSABIE的优化算法类似于之前所描述的ALE算法。区别在于WSABIE不使用任何先验信息，且其正则化值设置为0。

以上六种方法中，因为OVR、GLE和WSABIE不依赖于任何先验信息，所以它们不适用于零样本学习。

3. 零样本学习

作者在零样本学习条件下，对于AWA数据集，采用四十个类别用于训练，剩下十个类别用以评估。对于CUB数据集，作者150个类别用于训练，50个类别用以验证。

3.1 ALE输出编码的比较

作者比较了三种不同的输出编码：

- 连续编码
- 二进制(0,1)编码
- 二进制(-1,1)编码。

同时还比较了输出嵌入的两种归一化操作：

- 均值中心化
- L2归一化

结论：

- 连续编码比二进制编码的效果好的多。在AWA数据集上，连续编码的准确率比二进制编码正确率高了6.7%。在CUB数据集上，连续编码准确率比二进制编码高了7.3%。由此可证明连续编码所携带的类别与属性直接的关联信息更加丰富。
- 在连续编码的条件下，均值中心化对正确率的影响不太大，而L2归一化在AWA和CUB数据集上分别能提高2.4%和2.3%的准确率。

所以，在后面的实验中，作者总是使用连续编码和L2归一化。

3.2 学习算法的比较

作者比较了三个用于学习输入与输出之间的映射的目标函数，分别是：

- 岭回归(RR) 在AWA上正确率为44.5%。
- 结构化多分类支持向量机(SSVM) 在AWA上正确率为47.9。
- WSABIE的排名目标函数在AWA上正确率为48.5%。

由上可知RNK正确率最高，所以作者在后面都采用了RNK来学习模型的参数。

3.3 属性相关性

作者通过减少输出空间的维数，然后研究其对分类精度的影响，进而研究了输出空间中相关性。

作者分别研究了两种方式去减少输出维度:奇异值分解(SVD)和属性采样。

最终作者发现，属性直接存在这显著的相关性。而且虽然SVD不能保证选择总能选择信息最丰富的维度，但是SVD的性能优于属性随机采样。在属性随机采样中，属性的选择视乎更加影响输出嵌入的描述能力。

3.4 属性可解释性

在ALE中的W的每一列可以解释为属性分类器， $\theta(x) \cdot W$ 可以看作x的属性分数，但是作者没有像DAP一样直接优化属性分类器，所以可能丢失属性的可解释性。但是通过测量DAP与ALE的属性预测精度，发现ALE的属性精确度与DAP是比较接近的，所以ALE所学习出来的属性分类器依然是可解释的。

3.5 ALE、HLE和WLE的比较

在AWA上，ALE的准确率为48.5%，HLE的准确率为40.4%，WLE的准确率为32%。在CUB上，ALE的准确率为26.9%，HLE的准确率为18.5%，WLE的准确率为16.8%。WLE的准确率明显低于ALE与HLE的原因是，ALE与HLE主要依赖强监督，而WLE主要依赖的无监督学习。

3.6 与state-of-art相比较

目前在AWA上最佳的零样本识别正确率是Yu等人的48.3%,而测试的ALE能达到48.5%,经过ALE与HLE的预测分数融合后，准确率能够达到49.4%。由此可见作者提出的算法相对与目前最佳效果有一定的提升。

4. 少样本学习

实验设置:作者确定了一些感兴趣的类别(AWA中为10个,CUB中为50个)，作为评测类，然后从这些类别中再选取一半作为训练，另一半作为评测试样本。

作者将ALE与三个基线模型(OVR、GLE和WSABIE)进行了比较.其中作者主要分析了以下几个方面的问题:

- 有无参数共享对模型的影响
- 学习嵌入与固定嵌入所达到的不同效果
- 有无先验信息的影响

最后作者得出以下结论:

1. 对于有限的训练数据，GLE的性能优于其他所有方法。这表明在这种情况下随机嵌入是不合适的。
2. 在小训练集上，WSABIE和ALE的表现优于OVR或者GLE。这表明学习嵌入(WSABIE)或基于先验信息的嵌入(ALE)是有效的。
3. 在非常小的样本上，ALE的表现优于WSABIE，这表明先验信息在小样本学习上更加重要。

4. 当训练样本数量增加到比较大的时候，所有算法的准确率都差不多相同，这表明当训练数据比较丰富时，参数共享或者先验信息都不那么重要了。

5. 在完整数据集上进行测试

在本节中，作者保留了一半的数据用于训练，一半的数据用于测试。

5.1 输出编码的比较

作者和之前零样本学习时一样，比较了连续嵌入和二进制嵌入，以及均值中心化和L2归一化输出。然后作者得出了以下结论：

1. 均值中心化对正确率的提升没太大影响，而L2归一化能不断地提高准确率。
2. 在AWA和CUB数据集上，连续嵌入与二进制嵌入准确率相差不大，这表明当训练数据充足时，先验信息对标签嵌入的效果的影响较小。

5.2 不同嵌入方法之间的比较

作者比较了之前所述的不同方法。发现OVR在CUB上表现优于所有方法，GLE的AWA上也优于其他所有方法，这表明在训练数据集比较丰富的情况时，嵌入方法的质量已经不是太重要了。

六. 结论与分析

1. 在本文中，作者将基于属性的分类问题转换为了标签嵌入问题。
2. 作者提出的方法直接解决了图片分类问题，而没有引入属性分类等中间问题。
3. 作者的模型可以使用已标记好的训练数据来更新标签嵌入，即使用属性嵌入作为先验信息。
4. 标签嵌入不止可以使用属性信息，还可以使用文字语料中的层次结构信息，词嵌入信息等作为先验信息。
5. 连续编码可以提高零样本学习条件下的ALE的分类准确率，同时ALE学习出来的属性信息也具有可解释性。