Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly

一. 写作背景及目的

1. 背景

因为目前零样本学习在对缺少标记的训练样本分类方面起着越来越重要的作用，所以越来越多的零样本学习方法被提出来了。

2. 目的

分析一下目前的零样本学习现状并且定义一个通用的评价基准。

二. 核心问题

目前零样本学习的现状主要存在以下几点问题：

1. 目前零样本学习没有一个一致公认的评价基准，导致很难评价新提出来的方法在之前的方法基础上有了多大的进步。
2. 数据集拆分标准不一，导致之前公布的方法所到达的效果也没法直接进行比较。而且如果有人在测试类别上进行预训练会对结果产生比较严重的影响。

三. 现有状况

1. 常见的零样本学习方法类别

1. 线性方法(linear learning)
2. 非线性方法(nonlinear compatibility learning)
3. 独立分类器(independent classifier learning)

2. AWA1数据集

因为AWA1没有公开协议，目前不能直接获得原始图片，只能获得一些通过网络提取后的图像特征。

3. AWA2数据集

AWA2数据集具有公开协议协议，同时拥有和AWA1相同数量的图像类别和属性，且所有原始图片都是公开的，可以直接获取。

四. 相关研究

1. 早期(Two Stage)

1.1 通过属性(attribute)预测

步骤：

1. 预测出图像的一些属性
2. 搜寻具有这些相似的属性的类别

代表性方法:

1. Direct Attribute Prediction(DAP)
2. Indirect Attribute Prediction(IAP)

1.2 通过词向量(word vector)预测

步骤:

1. 预测已知类别的后验概率
2. 将图像特征投影到词向量空间(Question)

代表性方法:

1. IAP
2. CONSE

2. 近期的研究

大多数近期的零样本学习方法采用了直接学习一个从图像特征空间到语义空间的映射的方式。

2.1 SOC

SOC将图像特征映射进语义空间后，然后搜寻最相邻的类别向量。

2.2 ALE

ALE使用了排序损失函数，来学习图像与属性之间的双线性相容(compatibility)函数。

2.3 DeViSE

DeViSE采用了一个更有效的ranking loss来学习一个图像与语义空间线性映射，并且在大型的ImageNet数据集上进行了评估

2.4 SJE

SJE采用了SVM loss去学习双线性相容函数.

2.5 ESZSL

ESZSL采用了square loss去学习双线性相容函数.

2.6 SAE

SAE采用了一种语义自动编码器(semantic auto encoder),通过强制将图像特征投影到语义空间来对模型进行正则化。

2.7 LatEm

LatEm采用了多个线性映射,将SJE的双线性兼容性模型扩展到了分段线性映射.

2.8 CMT

CMT采用了带有隐藏层的神经网络,学习从图像特征到word2vec的非线性映射.

2.9 JLSE

JLSE将视觉特征和语义特征映射到两个独立的潜在空间中，并通过学习一个双线性相容函数来度量它们的相似性。

因为零样本学习在预测时只能使用没有参与训练的类，所以被批评其限制性比较强，所以，有人提出了广义的零样本学习设置，即将零样本学习任务推广到在测试时同时使用参与训练和未参与训练的类。

五. 作者所做的工作

1. 评估方法

零样本学习在训练时候，主要目的是通过训练时减少损失函数和正则项，然后学习出一个从输入到输出嵌入的映射。在测试时，零样本学习输入一个未训练过的类别的图片，然后将其进行正确分类，广义的零样本学习可以输入训练与未训练过类别的图片。

1.1 线性相容函数的学习

属性标签嵌入(ALE)、深度视觉语义嵌入(Designer)和结构化联合嵌入(SJE)都使用双线相容函数将视觉信息和辅助信息关联起来：

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

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$\psi(y)$之间的相容性（compatibility）。W是所要学习的视觉-语义映射矩阵。

1.2 非线性相容函数的学习

1. 潜在嵌入(Latem)

   Latem构造了一个分段线性相容函数:$F(x, y; W_i) = \max\theta(x)^T W_i \varphi(y)$,其中每个$W_i$对数据建立了不同的视觉特性，并且选择用于映射的矩阵是一个隐藏变量.Latem采用了的排序损失函数和随机梯度下降优化器。
2. 跨模态迁移(CMT)

   CMT首先将图像映射到了类名的语义空间，其中这个映射是含有tanh的非线性神经网络映射函数。最后再通过一种是否是新类的检测机制，可以将图像分配给看不见或看不见的类。

1.3 中间属性分类的研究

1. DAP首先为每个属性学习了一个概率分类器，并通过组合这些学习得到的属性分类器的分数来进行类别预测。
2. IAP首先通过预测每个训练类的概率，然后乘以类属性矩阵来间接估计图像的属性概率。

1.4 数据集相关

作者从大量的的零样本学习数据集中，我们选择了两个粗粒度数据集(一个小规模和一个中等规模)，两个中等规模的细粒度数据集，且都包含属性信息。还有一个是没有包含属性信息的大规模数据集。

1.5 含有属性信息的数据集

1. Attribute Pascal and Yahoo (aPY)是一个具有64个属性的小规模粗粒度数据集。在32个类中，作者选择了20个Pascal类用于训练（我们随机选择5个用于验证），12个Yahoo类用于测试。
2. 因为AWA1数据集的图像不公开，作者收集了和AWA1相同的50类别的的37,322幅图像作为AWA2数据集，同时AWA2与AWA1数据集相比，图像数量以及图像特征的分布。

1.6 大规模数据集ImageNet

作者还评估了所有方法在大规模数据集ImageNet上的性能。ImageNet具有21k个类别，但是每个类别所含有的图片数差别非常大。从ImageNet中可以提取出一个具有1K类，每个类包含大约1000个图像的平衡子集。作者没有采用像以前一样的把平衡子集拆分成测试集和训练集的方法，而是将平衡子集作为训练样本，然后把剩下的其他类别的样本作为测试数据进行测试。

2. 评估协议

2.1 图像和类的嵌入

1. 作者使用在ImageNet上使用1K平衡子集上预训练的ResNet-101模型提取特征，将输入图像进行前向计算后得到的2048维的池化单元作为图片嵌入.
2. 对于aPY，AWA1，AWA2，CUB和SUN数据集，作者使用了通过判断属性是否存在得出的二进制编码。而对于不含有属性信息的ImageNet数据集，作者使用了在的维基百科语料上训练过的Word2Vec来提取类嵌入信息。

2.2 数据集的拆分

在零样本学习中，在测试时的图像类别都应该是没有参与训练过的。但是用于提取他在的ResNet模型之前曾在ImageNet数据集上预训练过。同时，作者也主要到了有7个aPY测试类和6个AWA1测试类，都被包含在了ImageNet平衡子集中。所以作者提出了两种数据集划分方式，一种是和之前的一样的标准划分(SS)，还有一种是作者新提出来的建议的划分(PS)。在PS划分中，确保没有任何测试类出现在ImageNet 1K平衡子集中。同时作者也保持了SS和PS中的类别数相同。

2.3 评估标准

1. 如果采用对所有图像的TOP-1准确率进行评估，就可能会在样本比较多的类上得到的平均准确率的精度会比较高。然而，作者也希望在样本比较小的类别上得到的准确率也比较精准。所以,作者采用了以下方式来度量每个类的平均top-1准确率：

$$acc = \frac{1}{\|Y\|}\sum_{c=1}^{\|Y\|}\frac{\text{类别}c\text{中预测正确的数量}}{\text{类别}c\text{中总的数量}}$$

2. 在广义零样本学习设置中，评价时的搜索空间不仅限于测试类(Y^{ts})，还包括训练类(Y^{tr})。所以作者在计算了训练和测试类别中每个样本的的top-1平均准确率之后，作者又计算了一下训练的和测试时的准确率的调和平均值。如下图所示:

$$H = \frac{2*accY^{tr}*accY^{ts}}{accY^{tr}+accY^{ts}}$$

作者之所以没有采用算术平均值的原因是在算术平均值中，如果参与训练过的类精度更高，则会显著的影响整体结果。

六. 实验验证

1. 零样本学习实验

1.1 比较最先进模型

作者使用了使用原始论文中的公开特征和代码重新评估了所有方法,数据集划分此时采用标准划分方式(SS)。

1. 从结果中可以发现，作者复现的DAP，SYNC和SAE的结果几乎与其原始论文中的结果相同。
2. 同时作者发现使用手工提取的图像特征时，其准确率明显低于深层特征的结果。由此可见，改善视觉特征可以提高零样本学习的效果.

1.2 作者提议的分割方式(PS)

作者将所得到的建议的拆分(PS)结果与之前的标准拆分(SS)结果进行了比较。

1. 作者发现对于AWA1和AWA2，PS的结果明显低于SS。这是因为SS中AWA1和AWA2的大多数测试类都与ImageNet1K重叠。
2. 同时对于一些细颗粒度的数据集CUB和SUN，结果变化并不显著，因为在这种情况下，与ImageNet1K类别重叠也并不显著。
3. 同时作者还发现SYNC和SAE在SS上表现良好，此时SYNC能在SUN和APY上达到最佳性能，而SAE在SS划分模式下在AWA1和AWA2上的的最佳模型，而在PS中的性能要低得多，这表明它们在零样本学习任务中并没有达到很好的应用效果。

1.3 对方法排名进行可视化

作者将每个方法在第一至第十二位排列的次数作为其秩矩阵。然后，作者计算了每种方法的平均秩，并根据数据集的平均秩排序。作者发现：

1. 标准拆分(SS)上排名最高的方法是SYNC，但是在作者提议的拆分(PS)方式上，它下降到第七位。
2. ALE在SS上排名第二，在PS上排名第一。由此可得出结论：ALE方法似乎是对属性数据集上的零样本学习条件下最稳健的方法。
3. 在所有分类方法中，排名最高的三种方法是相容性学习方法，，而排名最低的三种方法是属性分类器学习或混合方法。因此可知，与学习独立属性分类器相比，相容性学习方法在零样本学习任务中的效果一直比较好。

1.4 在作者提出的**AWA2**数据集上的结果分析

AWA2与AWA1相比，具有相同的类和属性，但是包含不同的图像，每个图像都带有版权许可,可以公开访问并获取。

1. 在数据集采用标准划分(SS)的情况下，在大多数方法中，AWA1的结果与AWA2相近。
2. 在作者建议的数据集划分方式(PS)下的结果在AWA1和AWA2中也是一致的。对于12种方法中的8种，AWA1和AWA2的性能差异在2%以内。
3. 为了验证AWA2是AWA1的良好替代品，作者对12种方法进行了跨数据集评估。最后作者发现，所有在AWA1上训练的模型都能很好地推广到AWA2，反之亦然。
4. 同时作者发现与测试集相比，跨数据集的结果更依赖于训练集。最后作者通过实验对比表明如果测试集相同，但训练集不同，则结果有显着性差异。总之，训练集是衡量最终结果的重要指标。

1.5 ImageNet数据集上零样本学习结果

1. 在ImageNet等不含属性信息的大规模数据集上，使用Word2Vec提取词向量信息的条件下，SYNC方法取得了最佳性能。
2. 从所有方法的结果中可以看出，在图片数量最多的类别中的效果，要好于图片数量最少的类别，这表明在细粒度ImageNet子集上进行零样本学习是一项比较困难的任务。
3. 同时作者还发现一些测试集的自身性质，如被测试类别的类型，比分类的数量更重要。因此，测试集的选择在大规模数据集的零样本学习中显得比较重要。

2. 广义零样本学习的结果分析

作者在本节中使用了和上一节中相同的方法模型。经过实验，作者发现:

1. 在作者提议的数据集划分方式下，广义零样本学习的结果明显低于零样本学习结果.这是因为训练类别也被包括在了搜索空间中，对来自测试类别的图像进行分类时形成了干扰。
2. 相容性学习方法，例如ALE、Designer、SJE，在测试类别上表现良好。然而，学习独立属性或对象分类器(如DAP和Conse)的方法在训练类别上表现良好。

2.1 对方法排名进行可视化

1. 单纯评估测试类的TOP-1准确度而言，准确率最高的5种方法是ALE、Designer、sje、Latem、ESZSL。综合评估训练和测试类的调和均值时，性能最好的依次是ALE、Designer、Sje。
2. 经过对比，作者发现在评价广义零射学习时，不仅要优化测试类别的准确性，而且要对训练类别的准确性进行优化。
3. 作者还发现在广义零样本学习中，简单的新类别检测方案有助于提高结果正确率，因为与传统的零样本学习模型相比，所提出的新类别检测机制使用了更多的监督。

七. 结论

1. 在本文中，作者评估了大量的先进的零样本学习方法，最后发现，在统一的评估协议下，无论是零样本学习，还是广义的零样本学习，相容性学习模型的效果都要比学习独立的对象或属性分类器以及混合模型的效果更好。
2. 作者还发现一些标准的数据集分割方式中，测试时的一些类别可能会与预训练特征提取模型时所用到的图像类别产生重叠，从而影响对于零样本学习的准确率的评估的精度。同时作者提出了一种新的数据集分割方式，确保所有数据集中的测试类都不属于ImageNet1K。
3. 同时作者还引入了AWA2数据集。AWA2与AWA1在图像数量以及图像特征的分布，但是AWA2具有公开的版权许可，可供下载原始图片。同时实验表明，作者评估的12种方法在AWA2和AWA1上的表现类似。
4. 作者评估的12个模型的广义零射击学习精度明显低于它们的零射击学习精度，但不同模型的相对性能比较保持不变。
5. 作者还发现有些模型在测试数据全是未见过的类别时性能比较好，而有些模型在测试数据全是见过的类别时性能比较好。为此，作者提出了一个将见过的类和未见类的准确率的调和均值作为广义零样本学习环境下性能的统一度量的方案。调和均值评估方案在见过的和未见过的类的样本上都表现良好，这更接近于真实世界的设置。
6. 作者的以上工作广泛地评估了零样本学习的好方面和坏方面，同时消除了丑陋的方面。