尹浚宇 161130118

论文背景

卷积神经网络是现今视觉识别领域的前沿技术, 其在各种视觉识别任务中都有着优异的表现. 一般而言, 要使得卷积神经网络在应用中表现优异, 需要在训练阶段使用标记准确的大规模训练集. 换言之, 现有的卷积神经网络无法在训练集不充分或者标记有歧义的情况下学得一个令人满意的模型.

然而, 在视觉识别的很多领域中, 获取大规模标记准确的数据集是非常困难的. 首先, 在某些任务中, 就算是相关领域的专家也无法给出百分百正确的标记, 例如在语义分割中, 物体边缘像素的标记就存在着固有的歧义性. 其次, 获取完整且充分的数据是困难的, 例如要建立1-85岁的人脸数据库, 保证每个年纪都有着充分的数据, 这是难以实现的.

考察现在主流的标记处理方法(SLR, MLR, LS等), 可以发现它们都没有有效利用标签歧义性中蕴含的信息, 就算是一些最新发表的方法也只是在分类器学习阶段利用了相邻标记的相关性, 而在特征学习阶段, 则完全没有考虑利用标签歧义性.

综上所述, 我们有理由相信, 如果能充分利用标记二义性中潜藏的信息, 我们就可以在放松对数据集规模要求的同时, 在某些任务上取得比现有技术更好的表现. 该论文主要解决的问题就是如何利用标记二义性中潜藏的信息.

问题分析与建模

为了充分利用标记二义性中潜藏的信息, 作者的想法是对于某个样例, 根据其标签的真实值, 将其标签从单个值转换为一个离散的随机分布. 我认为这是一个非常自然的想法, 因为一个恰当的离散分布天然蕴含了标签二义性所包含的信息. 同时, 该想法等价于在训练数据总量不变的前提下增加了每个类别的数据量, 例如在年龄识别中, 真实值为25的图像对于识别年龄为24或26的图像也起到了一定的作用, 这样就可以放宽对训练集规模的要求.

同时, 作者的想法还包括将标记二义性同时应用在特征学习和分类器学习阶段. 我认为这也是一个很自然的想法, 因为在越多的环节利用标记二义性, 其中蕴含的信息的利用程度也就越高, 对结果的提升也就越大.

将上述想法形式化, 作者提出了DLDL方法. 总体而言, 该方法利用深度神经网络进行训练, 是一种端到端的学习方法. 其创新点在于输出不再是传统的单个类别输出, 而是一个离散的概率分布, 同时深度神经网络集特征学习和分类器学习于一体, 这说明该方法很好地综合了作者的想法. 使用文中的符号定义, 下面对该方法的细节进行说明.

首先, DLDL方法的输入空间为, 分别为输入图像的长,宽和通道数.然后DLDL直接从数据集学习后验概率, 其中是DLDL框架中的参数, 是一个类别分布向量, 是某个特定任务的标记集. 这里作者假设是完全的, 即其包含所有可能的类别取值, 该假设显然是合理的.

其次, 考虑如何修改神经网络来实现提出的想法, 包括如何将输出改变为概率分布, 在后向传播中使用什么样的损失函数, 框架中的参数如何更新, 当学习完毕后如何预测等. 给定一个样例, 假设是深度卷积神经网络最后一个全连接层的输出, 作者使用softmax函数将其转换为概率分布. 我们知道DLDL的目标是学习一个最优的, 使得在该参数下生成的分布最接近真实分布, 自然地, 我们需要能够度量两个分布之间相似性的损失函数. 这里作者采用KL散度作为分布相似性的度量, 于是有, 即损失函数可定义为. 作者使用随机梯度下降来更新参数, 忽略推导步骤, 有. 当训练完成后, 任何新样例的标记分布可由网络的一次前向运行得出. 特别地, 如果期望的输出结果是单个类别标记, 则DLDL输出, 满足. 如果期望输出是多个类别标记, DLDL输出, 是一个提前指定的阈值. 如果期望输出是一个实数, 则DLDL输出的期望, 即. 这意味着DLDL同时适用于分类和回归任务.

最后, 考虑如何为数据构建标记分布, 我们知道大部分数据集的标记都不是我们所期望的标记分布, 所以需要使用某些方法将其转换. 针对这一问题, 作者先提出了这些方法必须满足的原则: (1) 必须是合法的概率分布. (2) 中的分量应该有不同的取值, 即该分布应该集中在真实值附近而不是均匀分布. 之后, 作者分别针对年龄估计, 头部姿势估计, 多标记分类和语义分割提出了构建标记分布的方法.

对于年龄估计和头部姿势估计, 作者假设分布应该集中在真实值附近, 于是均使用正态分布来产生分布. 具体而言, 年龄估计的标记是单个值, 于是使用来产生分布, 头部姿势估计的标记包含了俯仰角和偏向角, 是一个向量, 于是使用来产生分布, 其中和分别是一维和高维正态分布的p.d.f.

对于多标记分类, 作者借用了PASCAL VOC数据集中的分组概念, 将标记分为Positive, Negative和Difficult三组, 其中Positive代表图像中包含该类物体, Negative反之, Difficult代表该物体可见但不易识别. 作者假设不同组标记概率不同, 并满足, 然后使用正则化产生分布, 其中根据属于的组分别取.

对于语义分割, 每个像素位置对应个标记, 表示位置为的像素, , 其中类别总数为, 图像背景用表示. 在FCN中, 位置的像素真实标记为意味着, . 为了产生分布, 该任务里等价于将标记的第三维转换为合理的概率分布, 作者使用了高斯核矩阵, 得到

其中, 是核矩阵大小, 和分别是填充大小和卷积步长. 然后使用来产生分布.

论文后面的内容包含了对神经网络中训练细节的说明(学习率, 不同层参数的选择等), 以及在上述四个任务中使用DLDL方法和其他方法的对比, 这些都不涉及DLDL的核心思想, 这里不再赘述. 值得一提地, DLDL在上述四个任务中均比其他方法取得了更好的效果, 通过论文的讨论部分可知, 其不仅可以在数据量小时有效避免了过拟合, 还有着更快的收敛速度和更好的鲁棒性.

问题拓展

考虑到KL散度具有不对称性, 不同的参数顺序可能会导致结果的不稳定, 有些情况下KL散度还会出现突变, 导致某点不可微, 这里我觉得可以使用Wessertein距离[1]来替代KL散度.

在生成标记分布的时候, 论文里的做法是人为根据数据集的特点进行设计定义的, 对于不同的应用场景, 需要设计不同的生成方法, 我觉得泛化性能不是很强. 经过搜索之后发现, 吴老师后面发表的论文[2]提出了解决方法, 具体想法是在反向传播中同时更新神经网络的参数和输入数据的标记.

事实上, 大部分多分类或者多标记的视觉识别任务都存在标记混淆和数据量不足的问题, 如果能为这些问题设计合理的标记分布的生成方法, DLDL技术就可以用在这些任务里.

参考资料

1. https://zhuanlan.zhihu.com/p/93853664
2. https://arxiv.org/pdf/1903.07788.pdf