《模式识别》课程论文作业

张逸凯 171840708 计算机系 本科生

本篇论文为 Deep Label Distribution Learning with Label Ambiguity, 其解决的**问题**为:

深度学习(例如卷积神经网络)的识别性能很大程度上依赖于一个较大的训练数据集, 并且需要被准确标记, 但是在某些领域无法满足这一要求:

- 比如年龄预测、头部姿态估计领域: 无法收集到完整和足够的训练数据, 不同类别上都分布足够数据较困难.
- 因为客观原因无法准确标记, 比如语义分割中靠近边界的像素难以标记, 年龄预测、头部姿态估计的标签在小范围内都可以认为是正确的.

该问题描述了深度学习模型卷积神经网络**过于依赖**大规模标注数据集, 这在几年后的 2020 年仍然存在并且是深度学习的**瓶颈**之一, 由此可以延伸出许多方法(例如迁移学习、Meta-Learning(将在最后一节讨论)), 该问题的本质为少样本学习并且减少过拟合、提高泛化性能. 这对于深度学习的发展**至关重要**. **现有方法**例如**基于标签平滑**(LS)或局限于标签之间的均匀分布, 或在特征学习阶段忽略了标签不确定性.

在下面的描述中, 沿用原论文的语言, 输入为X, 单一样本为 x_i, 标注的标签(真值)定义为 y, 标签取值域为 $\mathcal{Y} = \{l_1, l_2, \dots, l_C\}$.

为了**解决**上述问题, **关注到**数据集标签 y 之间存在**不确定性(label ambiguity)**, 提出了**两点**:

1. 标签到标签分布的转换:

即预先将训练样本的标签转换为标签所有取值域上的分布 $\mathbf{y} \in \mathbb{R}^{|\mathcal{Y}|}$, 捕捉样本标注的标签与其他可能"错误"标签之间的关系, 此转换的**理论基础**是合理、可解释的:

- 模型期望具有相似输出的输入图像之间具有很高的相关性, 比如年龄预测领域 32 岁和 33 岁图像之间的相关性应该强于 32 岁和 64 岁之间的相关性.
- 标签在小范围内都被认可, 比如年龄预测, 一般用"25 岁左右"这样的邻域来预测.
- 对提供给样本的标签只有部分信心, 比如识别任务中清晰出现却难以识别的对象.

此转换对于上述问题是**可行的**, **原因**如下:

- 总体来说每个类别关联的训练样本数量显著增加, 但实际并不增加训练样本总数.
- 减少过度学习, 带来训练后更强鲁棒性, 而且有效地降低了对大量训练图像的要求.

形式化层面, 出发点为尽量还原数据的真实分布, 可描述为:

- $\mathbf{y}$ 是一个概率分布, 即满足 $y_i \in [0,1], \sum_{i=1}^{C} y_i = 1$.
- 比较可能是输入图像真值标签的应该被赋予高概率.

2. DLDL(Deep Label Distribution Learning)方法:

在上述标签到标签分布的转换基础上, 作者提出 DLDL, 旨在**学习**输入 X 和 标签分布$\mathbf{y}$, 其任务是预测类似真值标签分布的 $\hat{\mathbf{y}} = p(\mathbf{y}|X;\theta)$. DLDL 在训练与测试时处理整个标签域的分布, 被迫学习**标签之间的不确定性(label ambiguity).**

DLDL 的处理该问题的**优势**为:

- 是一种端到端的学习框架, 应用于多种任务.
- 在特征学习和分类器学习中都利用了标签不确定性.

- 只有单一模型, 没有集成, 但是在年龄和头部姿态估计任务上取得比 SOTA 模型更好的性能.

DLDL 的**形式化**层面, 出发点即为找到真值分布和预测标签分布的**相似性度量**:

- 使用 Kullback-Leibler **(KL)散度**, 最优参数 θ^* 为:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_k y_k \ln \frac{y_k}{\hat{y}_k} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} - \sum_k y_k \ln \hat{y}_k$$

由此我们容易写出损失函数并应用链式法则**完成反向传播的推导**, 使用 **SGD** (Stochastic Gradient Descent)即可优化.

下面**举例**在特定领域标签分布的**构造**, 由**中心极限定理**, 我们很自然地选择正态分布:

- 年龄预测标签分布为: $y_j = \frac{p(l_j|\mu,\sigma)}{\sum_k p(l_k|\mu,\sigma)}$
- 头部姿态估计: 因为标签有两个维度: 俯仰角和偏航角, 所以标签取值域是二维的: $L = \{l_{jk}|j = 1,\cdots,n_1, k = 1,\cdots,n_2\}$, 其中$l_{jk}$是一对值, 其标签分布为: $y_{jk} = \frac{p(l_{jk})}{\sum_j \sum_k p(l_{jk})}$
- 多标签分类: 推广到较简单的形式, 即一个样本的多个标签都有三个层次: 已识别(Positive), 未识别(Negative), 清晰但难以识别(Difficult), 定义每个层次固定的概率且满足$p_P > p_D > p_N$, 含有某层次标签 $p(l_j)$ 即取上述三种概率之一, 其标签分布为: $y_j = \frac{p(l_j)}{\sum_k p(l_k)}$.
- 语义分割领域: (这里是比较难懂的地方) 维度更高(不止如上三个识别层次): y_{ijk}, 描述像素二维ij分量, 描述标签一维k分量. 考虑 $f_{K\times K}$ 的高斯核, 构造标签分布: $y_{ijk} = \frac{y''_{ijk}}{\sum_k y''_{ijk}}$, 其中:

$$y''_{ijk} = \sum_{i'=1}^{K} \sum_{j'=1}^{K} f_{i'j'} \times y'_{i'+(i-1)S-P, j'+(j-1)S-P, k}$$

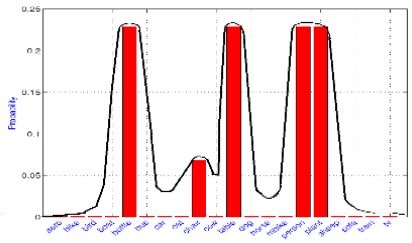
实验方面, DLDL 超越了年龄预测、头部姿态估计领域的 SOTA 模型, 在多标签分类和语义分割领域也有很好的表现.

从实验结果上看 DLDL 泛化性能得到提优的**可解释性研究**:

- 低维空间的**特征可视化**表明: 相比手工的特征 DLDL 嵌入后的结果有更好的语义聚类.
- DLDL 在标签一侧增强数据, 实验结果表明当数据集较小时这减少了过拟合风险.
- 分析实验结果还可以发现 DLDL 能**更快收敛**且具有**更强的鲁棒性**.

最后, 对论文的讨论:

- 论文的**关键之处**: 构造了一个真值意义下的后验概率分布进行端到端训练,仍是判别式模型, 由构造的后验概率分布强迫学习标签不确定性(label ambiguity), 来得到性能提优.
- **用想法 *A* 来代替论文中的想法:** 注意到原论文中多标签分类的标签分布是离散的存在零值的, 我们可以替换为单标签分类的推广: 设最多有C个标签, 如果某一样本真值为其中k个标签, 多标签分布对应着k个高斯混合的分布(类似 GMM), 其中每个高斯波峰的值为原论文中的 y_j, 如下图黑色部分:



这样做的**好处**是将原论文中离散的分布(并且存在零值)转化成了连续的, 关注到了样本中可能没有出现的标签. 对训练也是友好的. 但是在构造时可能较复杂.

实际上**我尝试了修改论文源码**并自己训练比较结果(但是 GPU 资源要求不够只好作罢), 点击这里是年龄预测的标签分布构造代码.

- **把论文中的方法向 *B* 问题拓展:**

1. 采用 Meta-Learning 构造标签分布:

由上讨论我们注意到构造的真值标签分布至关重要, 但是(高维)正态分布的方差为超参数(查阅源码验证点击这里论文中也有提及), Meta-Learning 即 learn to learn, 关注学习模型的 Learning Algorithm, 而不同的真值标签分布为不同的 Learning Algorithm, 并且 DLDL 是多任务的, 符合 Train Tasks 的划分. 请注意这里 Meta-Learning 关注的是构造什么样的标签分布, **在模型训练后**能得到较优的性能.

2. Few-Shot Learning 中可以加入 DLDL 作为性能提优的学习框架:

例如标签之间满足不确定性: 比如年龄估计的在小邻域内都被认可, 或者对于多个标签的预测只能给出信心值, 则可以用 DLDL 在后一阶段的学习进行优化.