基于训练数据集动态分割和分类采样的深度学习方法研究

杨鹏飞

摘 要

深度学习是由神经网络发展而来的一种实现机器学习的主要方法，在图像识别，自然语言处理，机器人控制等领域都有非常广泛的应用。深度学习方法需要从数据中进行训练学习，因此如何更好的使用数据来提高训练效率和准确率具有重要的研究意义。本文围绕这一课题得到了如下研究成果：

(1)提出了基于训练数据集渐进性分割的神经网络训练方法，其主要思想是：对神经网络模型进行多次迭代训练。每一次迭代训练都对训练数据集进行分割，让神经网络模型在分割得到的若干个子数据集上进行独立训练，根据这若干个在子数据集上训练后的神经网络模型在整个训练数据集上的误差表现，筛选出最优的即误差最小的神经网络模型，作为下一次迭代训练的初始网络。这里的渐进性分割是指数据集分割的过程是动态变化的，要求后一次分割要比前一次分割得到的子数据集个数少，尺寸大，即每个子数据集包含的数据样本更多。实验表明这种基于训练数据集动态分割的训练方法具有两个优点，一是较高的训练效率，因为每次分割得到的子数据集之间是相互独立的，针对每次数据集分割可以进行并行训练，起到加速训练的效果。二是这种训练数据集由小到大、由少到多的渐进性训练模式，符合人类对事物认识的变化规律，使得模型具有更好的训练准确率和测试准确率。

(2)提出了一种新的随机梯度算法(K-Mini-Batch SGD)，这种算法在传统的随机梯度算法(Mini-Batch SGD)训练基础上通过分类采样增加了样本数据的类别信息，这使得它在进行梯度估计时，得到的结果比传统的随机梯度算法更加精确，即所得的梯度估计能够更好的逼近整体样本数据集的梯度。由于它包含了每一类样本的梯度信息，所以在统计上具有更低的方差上界。通过理论分析，证明了新型随机梯度下降算法是收敛的，并且有更好的收敛效果。同时，通过合理的设置分类采样的参数，这种算法相比传统算法也没有增加计算复杂度。实验结果表明新提出的随机梯度算法是有效的并且准确率更高。

关键词：神经网络，训练数据集，渐进性分割，分类采样，随机梯度算法，准确率