自动化所提出一种基于哈希的二值网络训练方法

文章来源：自动化研究所 发布时间：2018-02-22 【字号： 小 中 大 】 我要分享

近年来，深度卷积神经网络已经深入了计算机视觉的各个任务中，并在图像识别、目标跟踪、语义分割等领域中取得了重大突破。在一些场景下，当前深度卷积网络性能已经足以部署到实际应用中，这同时也促进了人们将深度学习落地到更多的应用中。但深度卷积网络在实际部署时还面临着参数量和时间复杂度等两方面的问题，一方面是深度网络巨大的参数量会占用大量的硬盘存储和运行内存，这些硬件资源在一些移动和嵌入式设备中往往是很有限的；另一方面是深度网络的计算复杂度较高，使得网络推理速度很慢，同时会增加移动设备的电量消耗。

为了解决此类问题，研究者们提出了多种网络加速和压缩方法，其中的网络参数二值化是一种将网络参数表示为二值参数的方法。由于二值网络中参数只有+1和-1两种值，乘法运算就可以被加法运算替代。乘法运算比加法运算需要更多的硬件资源和计算周期，而使用加法运算替代乘法运算能够实现网络加速的目的。另一方面，原始网络参数的存储格式是32位浮点数，二值参数网络只使用1位来表示+1或者-1，达到了32倍的压缩目的。但是将参数从32位量化到1位会导致较大的量化损失，当前的二值网络训练方法往往会导致较大的网络精度下降，如何学习二值的网络参数同时又不带来较大的精度下降是一个问题。

近日，中国科学院自动化研究所研究员程健团队的胡庆浩等人提出了一种基于哈希的二值网络训练方法，揭示了保持内积哈希（Innerproduct Preserving Hashing）和二值权重网络之间的紧密关系，表明网络参数二值化本质上可以转化为哈希问题。

在该研究中，给定训练好的全精度浮点32位网络参数W，二值权重网络（BWN）的目的是学习二值网络参数B并维持原始网络精度。学习二值参数B的最朴素的方式就是最小化B与二值参数B之间的量化误差，但是这种量化误差和网络精度之间存在着一定的差距，最小化量化误差并不会直接提高网络精度，因为每一层的量化误差会逐层积累，而且量化误差会受到输入数据的增幅。

一种更好的学习二值参数B的方式是最小化内积相似性之差。假设网络某一层输入为X，X^TW是原始的内积相似性，则X^TB是量化之后的内积相似性，最小化X^TW与X^TB之间的误差可以学习到更好的二值参数B。从哈希的角度来讲，X^TW代表着数据在原始空间中的相似性或者近邻关系，X^TB则代表着数据投影到汉明空间之后的内积相似性。而哈希的作用就是将数据投影到汉明空间，且在汉明空间中保持数据在原始空间中的近邻关系。至此，学习二值参数B的问题就转化成了一个在内积相似性下的哈希问题，该哈希主要是将数据投影到汉明空间并保持其在原始空间中的内积相似性。

该研究首先在VGG9小网络上对方法进行验证，并且在AlexNet和ResNet-18上超过当前的二值权重网络。在ResNet-18上，该方法比当前最好方法的精度提高了3个百分点。

相关研究成果发表在AAAI 2018上。

论文链接

（责任编辑：程博）