



(12) 发明专利申请

(10) 申请公布号 CN 112189205 A

(43) 申请公布日 2021.01.05

(21) 申请号 201980031862.1

(22) 申请日 2019.02.27

(85) PCT国际申请进入国家阶段日
2020.11.12

(86) PCT国际申请的申请数据
PCT/CN2019/076374 2019.02.27

(87) PCT国际申请的公布数据
W02020/172829 ZH 2020.09.03

(71) 申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72) 发明人 隋志成 周力 赵磊 刘默翰
俞清华 蒋洪睿

(74) 专利代理机构 北京同达信恒知识产权代理有限公司 11291

代理人 李若兰

(51) Int.Cl.
G06N 3/04 (2006.01)

(54) 发明名称

一种神经网络模型处理方法及装置

(57) 摘要

一种神经网络模型处理方法及装置,在不降低神经网络模型的精度及有效性的情况下,压缩神经网络模型,该方法包括:训练得到第一低比特神经网络模型,该模型包括至少两个操作层,至少两个操作层包括第一操作层和第二操作层,每个操作层包括至少一个操作,用于所述操作的参数和/或数据的值用N比特表示,N为小于8的正整数,压缩该模型得到第二低比特神经网络模型,压缩后的模型包括至少一个操作层,至少一个操作层包括第三操作层,第三操作层等效于第一操作层和第二操作层,至少一个操作层中除第三操作层之外的操作层与至少两个操作层中除第一操作层和第二操作层之外的操作层相同。

