知识蒸馏在神经网络中的应用

李　锐，周　勇，牛小明

（中国兵器装备集团自动化研究所，四川 绵阳 621000）

摘　要　近年来，神经网络取得了飞速发展，在图像、语音、自然语言处理等领域获得了良好的应用成果。然而，主流深度学习网络模型存在计算复杂度高、对存储空间和带宽要求高等特点，决定了难以直接部署到众多计算资源和带宽受限的移动应用中。因此，如何将模型有效压缩且尽可能少地损失精度这一课题逐渐引起学者们的重视。本文主要介绍了知识蒸馏的基本思想和方法，并选取了手写体识别算法进行知识蒸馏并部署到了晟腾310平台，最后讨论了该领域目前面临的挑战以及可能的发展方向。

关键词　神经网络；知识蒸馏；手写体；硬件平台部署

中图分类号：TP3　　**文献标识码**：A　　**文章编号**：1007-0745(2023)01-0001-03

近年来，深度神经网络（Deep Neural Network，DNN）在智能机器人、汽车自动驾驶等领域获得了广泛的应用并取得了良好的应用成果。但性能优异的DNN往往具有网络结构复杂、节点数量巨大等特点。早在2012年ImageNet竞赛中获得冠军的AlexNet就已具有超过6千万的参数，且模型占据的内存高达241MB。随后涌现的优秀神经网络如ResNet、VGG、GoogLeNet、DenseNet等具有更加优异的性能，但随之而来的是更加庞大的网络模型、更加复杂的网络结构，所以模型运行对内存需求和算力需求逐渐增加。目前来看将DNN模型部署到一些存储和算力相对较低的硬件设备上依旧具有一定的难度。所以，如何在保证神经网络性能的同时尽可能降低网络模型的复杂度，从而使得优秀的网络模型能够运行在更广泛的硬件设备上是近年来学界的热门课题，该项课题技术的进步也对人工智能的广泛应用有着积极的意义。本文将介绍知识蒸馏的基本方法，并展示一种手写体识别模型的蒸馏以及硬件设备的部署。

1 知识蒸馏

知识蒸馏的过程涉及教师模型（Teacher Model）、学生模型（Student Model）这两个模型。教师模型选取大型神经网络，具有网络复杂度高、参数量巨大的特点，识别效果好但是不适合在低算力低内存设备中运行的模型，这类模型依赖大型服务器训练这种高性能硬件。学生模型复杂度低，但需要结构与教师模型相近，是适合在硬件资源有限的平台部署的一类模型。知识蒸馏的主要思想是以教师模型的高识别准确率经验去指导并训练学生模型，使得学生模型的识别准确率较传统训练方法大幅提升，从而达到精简模型，降低模型部署门槛的目的。

Hinton等人[1]在2015年首次提出知识蒸馏（Knowledge Distillation，KD）这一网络压缩技术。在现有的神经网络训练方法的基础上，区别于传统的硬标签（hard-taget）只能给出唯一的识别结果，Hinton等人提出了软目标（soft-target），能够给出分类结果属于某一种类别的概率，这一参数用于计算出总损失函数用以指导训练学生模型。软目标（soft-target）的计算结果q_i可从以下函数计算得出。

$$q_i=\frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_i}{T}\right)}$$

与传统的softmax函数不同，这里引入了蒸馏温度T这一参数，公式中的z_i是输入数据为第i类结果的概率。在训练的过程中需要计算蒸馏损失（Distillation Loss）以及学生模型损失（Student Loss）。其中蒸馏损失是在选定的蒸馏温度下分别训练教师模型和学生模型后，通过交叉熵损失函数[2]计算出的结果；学生模型损失是学生模型在选定蒸馏温度T为1训练后与已知的硬标签（hard-taget）对比计算出的结果。结合以上两个损失参数可计算出一个新的损失函数，用该函数对学生模型进行反向传播。整个蒸馏过程如图1所示。

知识蒸馏压缩效果的评价指标包含以下三种：（1）学生网络中模型参数较教师模型的降低比率，这一指标直接体现知识训练对网络的压缩率；（2）通过知识蒸馏训练后学生模型识别效率与教师模型之间的差距；这一指标能够直观体现压缩后网络的精度损失量；（3）通过知识蒸馏这一训练的学生模型与常规训练后的学

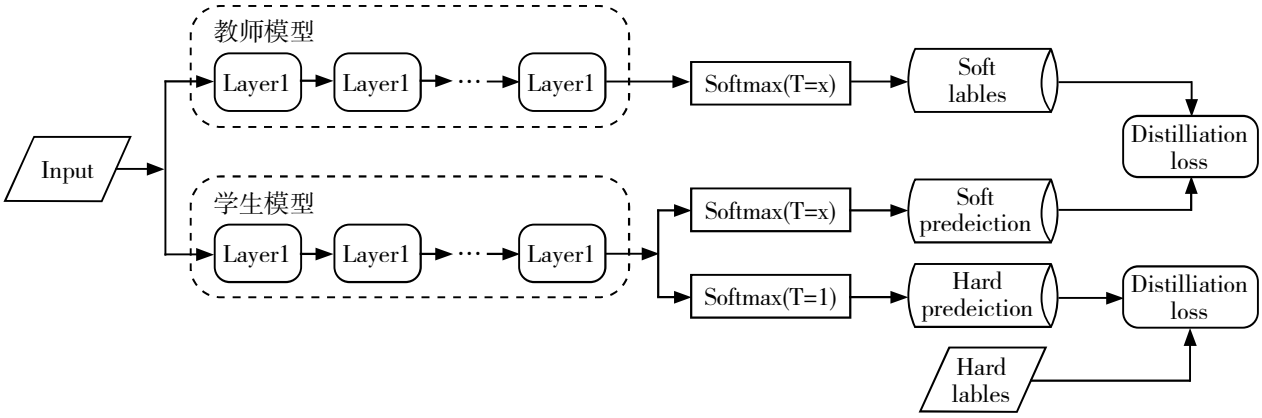

图 1 知识蒸馏基本流程

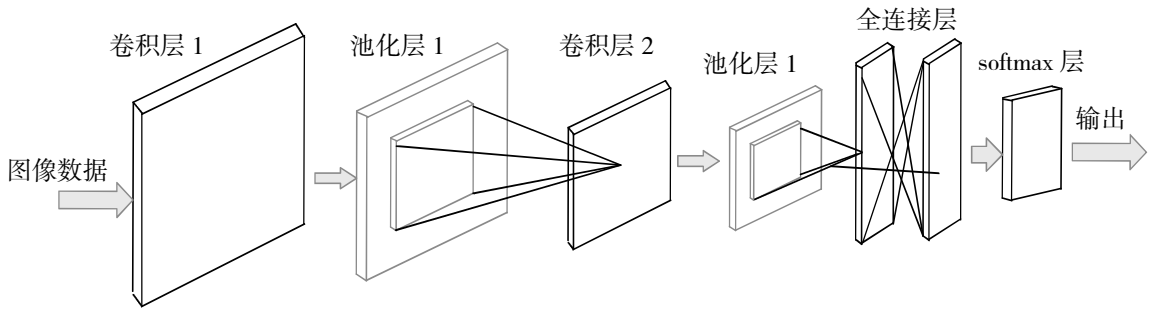

图 2 卷积神经网络构成图

生模型性能之间的差异。这一标准能够体现在一定压缩率（本次蒸馏训练）的情况下训练学习过程的有效率。

通过知识蒸馏的过程获得的学生模型，与教师模型相比复杂度明显降低，模型性能损失可控在一定范围，对低算力硬件平台友好。但是由于知识蒸馏对 softmax 函数的改造以及软标签（soft-target）的特性，目前知识蒸馏在分类任务中能够取得较好的效果，但在复杂的识别类任务中还有很强的局限性。并且学生模型的性能提升效果严重依赖一个性能良好且适用于该学生模型的教师模型进行训练，所以知识蒸馏在实际应用中还有很多需要发展的方向。

2 手写体分类的蒸馏与部署

2.1 手写体分类算法

字符识别一直是图像分类领域的一项重要应用，在签字文件、金融票据、邮件信息等方面的手写信息录入有着良好的应用前景。特别是在实时应用场景中，手写字母识别可以被认为是一种特殊的人机交互方式，具有很大的价值。本文中构建的手写体卷积神经网络构成如下。

2.1.1 输入层

输入层是神经网络数据预处理层，当网络的输入数据为图像数据时，图像信息经过输入层处理后会将转换为一个三维矩阵。该矩阵第一个维度是输入图像的高度，第二个维度是图像的宽度，第三个维度是 RGB 数据通道。由于不同的图像可能由不同的色彩分量主导，从而使得权值的主导权不断变化导致收敛速度变慢，故需要对 RGB 分量做标准化处理。

2.1.2 卷积层

卷积层的主要作用是对输入层预处理过的图像数据进行特征提取。具体规则是使用较小的矩阵对输入层得到的三位矩阵进行卷积运算，这里使用到的小矩阵又称做卷积核或滤波器，需要注意卷积核的深度需要与图像数据矩阵的深度一致。选取不同卷积核对应提取不同的特征，一般卷积核不宜过大。

2.1.3 池化层

池化层实现的效果类似于图片压缩，可以在保留有效特征值的同时降低参数复杂度。最大池化将矩阵划分为相同大小的区域，并在每个区域内选取最大的特征值，该过程忽略了提取的特征值的精确位置，而保留了特征值本生及其相对位置。通过该步骤网络中的参数量会迅速下降，不但提升了后级的计算速度，也可防止过拟合，提高网络模型的鲁棒性[3]。

2.1.4 全连接层

全连接层就是一个分类器，经过网络中输入层、

卷积层、池化层对数据的运算，可以将输入数据映射到特征空间，但是并未与样本标签产生关联。通过全连接层后，可直接将特征数据映射到样本标签，全连接层可直接由卷积操作实现。

2.1.5 Softmax 层

Softmax 层可以将全连接层算出的数值向量归一化为得到概率分布向量，各类标签的概率之和为 1。

2.2 蒸馏过程

1. 首先使用 MINST[4] 数据集训练每个隐藏层具有 2000 个神经元基于 Pytorch 搭建的教师模型，设置的训练 20 轮，得到在数据集上分类准确率为 98.32% 的 onnx 模型。

2. 构建一个每个隐藏层仅有 50 个神经元的学生网络。

3. 使用蒸馏温度 T=6 训练教师网络（此过程切断反向传播）得到 soft target1，训练学生网络得到 soft target2，结合以上两个参数使用交叉熵损失函数得到蒸馏损失（Distillation Loss），即学生模型与教师模型之间的误差。

4. 在蒸馏温度 T=1 的情况下对训练学生网络得到 soft target3，通过学生网络的训练情况与正确的标签结果进行对比，得到学生模型损失（Student Loss），即学生模型与正确结果之间的误差。

5. 通过以上步骤获得的蒸馏损失与学生模型损失利用以下公式得到学生模型反向传播时使用的损失函数 L。通过损失函数 L 对学生模型进行反向传播，即可完成对学生模型的知识蒸馏过程。

$$L = (1-\gamma)T^2 L_{dis} + \gamma L_{stu}$$

由于 softmax 函数中引入了蒸馏温度 T，反向传播过程中梯度受蒸馏影响变为原来的 $\frac{1}{T^2}$，为恢复梯度尺度，使其与真实标签对应的交叉熵的尺度一致，需要在计算交叉熵时乘以 T^2。

2.3 晟腾 310 平台的部署及运行结果

本文的硬件部署平台选取华为 Atlas 200DK[5]，它是基于升腾 310 人工智能芯片的一个开发板产品。该开发板以 Atlas 200AI 加速模块为核心，可以适配种类众多的机器学习模型对图片视频信息进行分类推理，该开发套件有良好的社区生态以及完备的技术指导资料，可快速搭建环境并进行开发，且支持的机器学习模型也较为广泛。

首先，准备好一张容量大于 32GB 的内存卡，一台有 Ubuntu18 系统并配置好交叉编译环境的 PC 机；使用读卡器连接内存卡与该 PC 机，使用 weget 获取制卡脚本 make_sdcard.py 和 make_sd_card.sh，并使用脚本制卡；制卡成功后使用 ssh 连接到开发板，并用 pip 安装 attrs、numpy、decorator、sympy 等相关依赖；之后安装 Ascend-cann-toolkit。至此开发板相关环境搭建完毕。

下面进行模型部署，模型部署的关键在模型转换，有 Mind Studio[6] 开发工具平台、ATC 命令这两种方法，由于本文中已配置好 CANN 环境和 ATC 工具，可直接将上文中训练好的上传至开发套件，并使用 ATC 命令将 onnx 模型转化为晟腾 310 支持的 om 离线模型。

蒸馏结果及结论：对于未蒸馏的学生模型，其在用 MINST 数据集训上的准确率为 89.68%。而蒸馏后的学生模型在相同数据集上的准确率提升至 91.63%。由此可以较为明显地看出蒸馏对准确率的提升。

3 总结与展望

随着人工智能的发展，神经网络会被更多地部署到各种硬件平台中，其中就包括大量资源有限的硬件，所以未来一定会对降低神经网络复杂度及降低对硬件资源的需求设计产生巨大的需求，知识蒸馏技术作为一种高效的网络压缩技术，在未来一定会获得更多的应用并获得长足的发展。知识蒸馏的效果很大程度上取决于学生模型能够从教师模型那里学到多少知识，即提取知识的效率。目前提取效率还有很大的提升空间。未来，如何让学生模型更高效率地从教师模型中提取知识必然成为知识蒸馏的发展方向。随着神经网络规模的不断增大，未来知识蒸馏将越来越多地和其他网络压缩方法结合交叉使用，例如剪枝、量化等方法。随着具有 AI 算力的硬件设备不断发展，未来人工智能算法高效的硬件部署也将成为重要课题。

参考文献：

[1] Hinton G,Vinyals O,Dean J.Distilling the Knowledge ina Neural Network[J].Computer Science,2015, 14(07):38-39.

[2] 崔义新 . 基于交叉熵的随机赋权网络 [D]. 保定 : 河北大学 ,2017.

[3] 廖明哲 . 基于深度学习的遥感图像匹配方法研究 [D]. 武汉 : 武汉科技大学 ,2020.

[4] Deng L.The mnist database of handwritten digit images for machine learning research [best of the web][J]. IEEE signal processing magazine,2012,29(06):141-142.

[5] 华为云 .Atlas 200 DK 开发者套件产品介绍 [DB/OL]. https://support.huaweicloud.com/ productdesc-A200dk_3000/atlas200_DK_pdes_19_0007.html,[2021-03-16].

[6] 华为云 .MindStudio 版本 :2.0.0(beta1) 用户指南 [DB/OL].https://support.huaweicloud. com/usermanual-mindstudioc73/atlasmindstudio_02_0004.html.[2021-03-16].