1. Introduction

最近几年，CNN已经成为了计算机视觉任务的主要技术手段，在图像分类、目标检测、深度估计、语义分割等方向都大放异彩。尤其是自从AlexNet一举夺得ILSVRC 2012 ImageNet图像分类竞赛的冠军后，深度神经网络的热潮便席卷了整个计算机视觉领域。自此深度模型火速替代了传统人工设计特征和分类器。各种深度模型不仅提供了一种端到端的处理方法，还大幅度地刷新了各个任务的精度，甚者已经超越了人类知识的准确度。目前为止，在图像分类模型的准确率已经达到95%以上（top5）。

In recent years, CNN has become the main technical means of computer vision tasks, and it has been splendid in the direction of image classification, target detection, depth estimation, and semantic segmentation. In particular, since AlexNet won the ILSVRC 2012 ImageNet Image Classification Competition in one fell swoop, the deep neural network boom has swept the entire field of computer vision. Depth model swiftness replaces traditional manual design features and classifiers. Various depth models not only provide an end-to-end processing method, but also significantly refresh the accuracy of each task, and even beyond the accuracy of human knowledge. So far, the accuracy of the image classification model has reached more than 95% (top5).

然而，深度模型优异的表现背后是强大的计算需求：在不断刷新任务精度极限的同时，其深度和尺寸也在成倍增长。为了保留用户隐私并减少用户感知的查询时间需要将这些深度神经网络迁移到各种新平台应用，其中包括移动平台，自治系统和智能设备等。因此随之而来的问题之一是：如此巨大的模型只能在有限的平台下使用，根本无法移植到移动端和嵌入式芯片当中，而且较高的带宽占用也让很多用户望而生畏；另一方面，大尺寸的模型也对设备功耗和运行速度带来了巨大的挑战。所以深度模型目前面临着在终端部署和低延迟需求场景下难以应用的问题。

However, behind the excellent performance of the depth model is a powerful computing demand: while constantly refreshing the limits of the accuracy of the task, its depth and size are also exponentially increasing. In order to preserve user privacy and reduce user-aware query time, these deep neural networks need to be migrated to various new platform applications, including mobile platforms, autonomous systems, and smart devices. Therefore, one of the following problems is that such a huge model can only be used on a limited platform, it cannot be ported to mobile terminals and embedded chips at all, and the high bandwidth consumption is also daunting to many users. On the other hand, large-scale models also pose enormous challenges to equipment power consumption and operating speed. Therefore, the depth model currently faces problems that are difficult to apply in the context of terminal deployment and low-latency requirements.

因此，模型压缩是让深度模型真正应用在移动端不可缺少的一步，近年来，该领域实现了极大的发展。目前主流的压缩方法包括了参数修剪和共享、低秩分解、知识提纯等。基于参数修剪和共享主要的工作包括两个方面：权值量化和权值修剪，前者关注减小模型大小，后者则关注于探索模型参数中冗余的部分，并尝试去除冗余和不重要的参数。在量化部分，Gong等人通过对参数值使用K均值的方法进行量化，。Han将权值共享和量化、编码等方式运用在模型压缩上。基于模型裁剪的方法很多，其思路源头都是来自于Oracle pruning 的方法，即挑选出模型中不重要的参数，将其剔除而不会对模型的效果造成太大的影响。基于低秩分解（Low-rank factorization）[12, 13]技术的方法使用矩阵分解以估计深层网络中最具信息量的参数，然后通过将一个大矩阵分解为多个小矩阵进行计算。知识提纯（Knowledge Distillation, KD）[14, 15]则学习了一个精炼模型，即训练一个更加紧凑的神经网络以再现大型网络的输出结果。KD的基本思想是通过软Softmax 学习[16]教师输出的类别分布而将大型教师模型（Teacher Model）的知识提纯为较小的模型。

Therefore, model compression is an indispensable step for the depth model to be truly applied to the mobile terminal. In recent years, this field has achieved great development. The current mainstream compression methods include parameter pruning and sharing, low rank decomposition, and knowledge purification. The current mainstream compression methods include parameter pruning and sharing, low rank decomposition, Knowledge Distillation. The main tasks based on parameter pruning and sharing include two aspects: weight quantification and weight pruning. The former focuses on reducing the size of the model, while the latter focuses on exploring redundant parts of the model parameters and And try to remove redundant and unimportant parameters. In the quantification section, Gong et al. quantified the K-means using parameter values. Han applies weight sharing and quantification and coding to model compression. There are many methods based on model tailoring. The source of their ideas are all derived from the Oracle pruning method. That is, picking out the unimportant parameters in the model and removing them will not cause too much impact on the model's effect. The low-rank decomposition uses matrix decomposition to estimate the most informative parameters in the deep-layer network, and then uses a large matrix to be decomposed into multiple small matrices for calculation, thereby reducing the amount of computation. Based on the Knowledge Distillation approach, it belongs to migration learning. Migration learning is the migration of the performance of one model to another. The teacher network is often a more complex network with very good performance and generalization capabilities. This network can be used as A soft target to guide another simpler student network to learn, making a student model that is simpler and less parametrically computational, can also have similar performance to the teacher network, and can be considered as a model compression approach.

因此，模型压缩是让深度模型真正应用在移动端不可缺少的一步，近年来，该领域实现了极大的发展。目前主流的压缩方法包括了参数修剪和共享、低秩分解、知识提纯等。关于深度模型压缩的理论研究国内外已经有所基础，但相对于其他任务（如计算机视觉，语音识别等），该领域目前还处于一个发展启蒙阶段，尤其是深度学习模型在真正移动端的部署还是很少。近两年来，相关研究工作也在逐步开展，而要想顺利的进行移动端的部署，一方面需要分析不同深度模型本身的性能，一方面模型的压缩也是非常必要的研究点。模型压缩主要是围绕减小模型尺寸大小和推理时间进行实验的。通过量化共享权值可以对模型的尺寸起到很好的压缩作用，通过修剪权重可以进一步加速模型的推理时间，但是这些方法主要还是停留在实验结论上并没有真正在移动端或者嵌入式系统上部署模型并实践效果，所以真正在嵌入式平台上部署模型后进行性能测试以及压缩实验对于进一步检验模型在移动端或者嵌入式系统上部署可行性是必要的一步，也是本文的工作重点。

In summary, the theoretical research on the depth model compression has been established at home and abroad, but compared to other tasks (such as computer vision, speech recognition, etc.), this field is still in an enlightening phase of development, especially in the depth learning model. The real mobile deployment is still very few. In the past two years, related research work has been gradually carried out. To successfully deploy mobile terminals, it is necessary to analyze the performance of different depth models on the one hand. On the other hand, the compression of models is also a necessary research point. Model compression is mainly about reducing the model size and reasoning time for experiments. By quantifying the shared weights, the size of the model can be compressed very well. By pruning the weights, the reasoning time of the model can be further accelerated. However, these methods mainly stay on the experimental conclusion and are not really on the mobile or embedded system. Deploying the model and putting it into practice, performance testing and compression experiments after actually deploying the model on an embedded platform is a necessary step to further verify the feasibility of the model's deployment on the mobile or embedded system, and is also the focus of this article.

为了更好的将模型部署到移动端或者嵌入式系统上使用，我们的工作进行两方面的工作研究：

（1）对目前最经典使用广泛的深度学习模型在JETSON TX2嵌入式系统上进行性能的测试和分析，包括推理时间，模型大小，各操作数的时间花费，进一步分析这些模型在嵌入式系统下的具体性能表现，相关研究工作目前还没有出现过，这也是深度学习模型在移动平台的移植的第一步。

（2）研究基于深度学习的模型压缩技术，实现了对几种经典网络的模型压缩，对比不同压缩方法下模型的负载特征，主要包括模型压缩前后的推理时间变化，模型大小变化以及准确率的的的损失等，从而得到不同压缩方法对不同模型的适用效果不同，分析不同模型适合的压缩方法，使得不同模型在移动端的使用更具可能。

In order to better deploy the model to a mobile or embedded system, This paper makes the following contributions:

(1) Performance testing and analysis of the most widely used deep learning models in embedded systems, including inference time, model size, and time spent on each operand, and further analysis of these models in embedded systems Performance performance, related research work has not yet appeared, which is the first step in the migration of the deep learning model in the mobile platform.

(2) Study the model compression technology based on deep learning, realize the model compression of several classical networks, compare the load characteristics of the model under different compression methods, mainly including the reasoning time change before and after the model compression, the model size change and the accuracy rate The loss, etc., resulting in different compression methods for different models of different applications of the effect of analysis of different models suitable for the compression method, making the use of different models in the mobile terminal is more likely.