Theme: （技术领域）

Computer vision

Topic: （需求名称）

软硬一体系统的异构任务调度与加速算法研究

A heterogeneous task scheduling and deep model acceleration research in software-hardware integrated system

Background

软硬一体方案是将端侧方案与阿里云现有的云边一体方案结合，形成一个云、边、端协同的多层架构。这种架构特别适合城市大脑中的复杂的、层次化的业务。不同层次的业务以一种最优的方案映射到云、边、端，形成一个统一的解决方案，达到降低运营成本、减少能源消耗、提高用户体验等作用。我们希望通过多目标优化方法实现云边端一体系统的任务调度，通过模型压缩算法实现推理加速，降低硬件成本，通过在线学习算法降低算法的适配成本。

The software-hardware integrated solution combines the end-side solution with Alibaba Cloud's existing cloud-side integrated solution to form a multi-layered architecture of cloud, edge, and end collaboration. This architecture is particularly suitable for the complex and hierarchical business in the city brain solutions. Different levels of business are mapped to the cloud side, the edge side, or the end side with an optimal solution to form a unified solution to reduce operating costs, reduce energy consumption, and improve user experience. We hope to achieve task scheduling of the integrated system with multi-objective optimization methods, achieve inference acceleration through model compression algorithms to reduce hardware costs, and reduce algorithm adaptation costs through online learning algorithms.

云边端一体系统通常由多层次的设备组成。计算设备具有异构性，其计算能力和能源功耗各不相同，与此同时，设备之间的网络连接也存在带宽和延时的多样性。一个复杂的计算任务需要多个不同层次的设备协作完成，如何对任务进行建模并调度到多层次的设备上是我们需要考虑的问题。为了选择最优的调度策略，我们可以定义算力、延时、带宽、能源消耗等多维度的度量方式。然而现实的复杂性通常要求我们在多个维度同时达到最优，为此我们需要

研究基于多目标优化的云边端一体系统任务调度算法。

The integrated system usually consists of heterogeneous devices. The computing power and energy consumption of heterogeneous computing devices are different. At the same time, the network connection in the devices also has diverse bandwidth and delay. A complex computing task requires multiple devices at different levels to cooperate. We hope to solve the problem that dispatching a task to multi-level devices. In order to select the optimal scheduling strategy, we can define multi-dimensional measurement methods such as computing power, delay, bandwidth, and energy consumption. However, the complexity of reality usually requires us to achieve the optimization in multiple dimensions at the same time. For this reason, we need to study the task scheduling algorithm of the software-hardware integrated system based on multi-objective optimization.

深度学习模型需要大量算力、内存。在软硬一体系统中，执行实时推断、在设备端运行模型、在计算资源有限的情况下运行模型，大型的深度模型将受到挑战。模型压缩正是提高推断效率、降低运行时内存占用的重要方法。如何通过对模型进行量化、剪枝、蒸馏、神经架构搜索，高效的生成规模更小、内存利用率更高、能耗更低、推断速度更快、推断准确率损失最小的模型，在软硬一体解决方案中是一个重要的课题。

Deep learning models require a lot of computing power and memory comsumption. In a software-hardware integrated system, large-scale deep models will be challenged when they are performing real-time inference, running models on the device side, and running models with limited computing resources. Model compression is an important method to improve inference efficiency and reduce runtime memory usage. How to efficiently generate models with smaller scale, higher memory utilization, lower energy consumption, faster inference speed, and minimal loss of inference accuracy through quantification, pruning, distillation, and neural architecture search on the model is an important topic in the integrated solution.

端侧部署的深度学习模型，一般需要以本地样本进行训练，产品进行大规模推广会急剧提升适配成本。如何通过在线自学习的方法降低模型适配的成本，是软硬一体解决方案中的重要课题。我们希望通过在线自适应的样本收集、模型训练算法，降低本地化部署的人工成本。

The deep learning models deployed on the end-side generally need to be trained with local samples. As products are broadly promoted, adaptation costs will increase dramatically. How to reduce the cost of model local adaptation through online self-learning is an important topic in the software-hardware integrated solution. We hope to reduce the labor cost of localized deployment through online adaptive sample collection and model training algorithms.

端上目标检测：针对端上视频中目标检测的特征融合算法

随着边缘端芯片的发展，使得模型在边缘端的推理成为可能。随着应用的需要，越来越多的场景需要在边缘端上进行推理，以满足应用在推理速度和网络带宽的要求。尽管如此，边缘端的推理仍然需要在速度和精度之间做出平衡。

With the development of edge chips, it is possible to make model inferences at the edge. With the needs of applications, more and more scenarios require inference on the edge to meet the application's requirements for inference speed and network bandwidth. Nevertheless, inference at the edge still requires a balance between speed and accuracy.

有许多的工作将一个已有的模型应用到边缘端进行在线推断，如通过使用模型压缩、模型量化，将已有的大网络如 resnet50、vgg 等进行压缩和量化，使其可以在边缘设备上获得较高的计算速度；另一种是直接设计一种计算量较小的模型，如 mobilenetv1～v4，shufflenet v1～v2，mnasnet 等等。

There are many works that apply an existing model to the edge for online inference. For example, by using model compression and model quantization, the existing large networks such as resnet50, vgg, etc. are compressed and quantized, so that they can be used on edge devices Obtain a higher calculation speed; the other is to directly design a model with a smaller amount of calculation, such as mobilenetv1～v4, shufflenet v1～v2, mnasnet and so on.

然而，这两种方法均会降低模型的精度。但是实际上，我们可以充分利用时域信息，并借助云端的资源，有望在算力受限的设备上能得到较高的精度。例如，我们可以使用云端低频检测的方式，通过将边缘端的图像以一个较低的频率发送到云端，云端使用一个精度高的模型计算得到大尺寸的特征，端侧将自身的特征和云端的大尺寸特征进行融合，这样既利用了云端推理无法实时但精度高的特点，又利用了端侧计算能力受限但速度较快的优势，通过特征融合的方式获得一个推理速度和精度都较高的模型。
However, both of these methods will reduce the accuracy of the model. But in fact, we can make full use of time domain information, and with the help of cloud resources, it is expected that higher accuracy can be obtained on devices with limited computing power. For example, we can use cloud low-frequency detection. By sending the edge image to the cloud at a lower frequency, the cloud uses a high-precision model to calculate the large-size features, and the end-side compares its own features with the cloud's large size. The dimensional feature is fused, which not only takes advantage of the fact that cloud reasoning cannot be real-time but high precision, but also takes advantage of the end-to-side computing power but is faster. Through feature fusion, a high inference speed and accuracy is obtained. model.

总之，我们希望实现云边端一体系统任务调度算法；通过模型压缩算法实现推理加速，降低硬件成本；通过在线学习算法降低算法的适配成本；通过云边端特征融合，提升算法效果。
In short, we hope to realize the task scheduling algorithm of the cloud-side-end integrated system; realize inference acceleration through model compression algorithm and reduce hardware cost; reduce the adaptation cost of the algorithm through online learning algorithm; and improve the algorithm performance through cloud-side feature fusion.

Target

1、一种多目标优化的云边端一体系统任务调度算法

2、一种深度神经网络推理加速方法

3、一种在线学习降低适配成本的方法

4、一种云边端特征融合提升算法效果的方法

1、 A multi-objective optimized task scheduling algorithm for cloud-side-end integrated system

2、 A Method for Accelerating Deep Neural Network inference

3、 A Method of Online Learning to Reduce Adaptation Cost

4、A method of cloud edge feature fusion to improve algorithm performance

Related Research Topics

1. 神经网络架构搜索
2. 在线目标检测与跟踪算法

1.Neural network architecture search

2.Online learning for object detection and tracking