目录

产品概述

金山云人工智能开发平台KPL（Kingsoft Cloud Power Learning，以下简称KPL），是金山云推出的面向AI开发和应用的一站式人工智能开发生产平台，实现了对数据、算法、算力资源进行统一调度管理，构建了完善的开发软件栈，支持数据处理、算法开发、模型训练、模型部署的全流程业务，并提供了大规模分布式训练、自动化模型生成等功能，满足不同开发层次的需要，有效提高计算资源利用率，提升AI开发效率，实现系统的平滑、稳定、可靠运行。 使用KPL，您可以在不关心算法细节的情况下，只需要提供数据以模板化的方式进行简单高效的算法迭代训练，落地算法应用，也可以使用Jupyter Notebook的方式进行算法研发等。

产品功能

（1）数据上传与管理。KPL平台内数据均使用KPL Dataset 格式。KPL Dataset是专门为解决AI大数据存储、数据高效预处理和用于大规模算法训练设计的数据存储格式，与TensorFlow的TFRecords文件性质类似。KPL Dataset读写简单高效，在KPL平台的所有数据处理和训练过程中都以该格式文件为基础进行了大量优化，以及在前端页面上专门对该格式数据进行适配直接渲染数据内容。

（2）开发环境快速部署。AI的算法开发需要复杂的GPU计算集群，涉及各种不同厂家不同型号的设备，且开发环境多样，部署十分复杂且耗时耗力，成本很高，KDL平台实现对计算资源统一智能分配管理，实现GPU用户配额和限制策略，保证计算资源能根据开发人员的需求进行合理的调度；计算资源按需申请，随用随到，支持GPU共享模式，提升集群整体利用率，通过自定义镜像等功能，实现开发环境的快速部署。

（3）自定义工作流模板：支持基于工作流自定义任务运行模板，用户可根据需求对数据集、预训练模型、数据预处理、训练、测试、批量推理、编译SDK进行自由排列组合，使用连线确定其关联关系和运行顺序，每个运行节点可单独配置参数，多个节点构成运行流图，可以保存作为某一场景下的任务模板。

（4）自定义算法运行：支持主流编程语言编写的任意AI算法。自定义的算法可以安装到平台中运行，平台不对算法类型、算法应用场景等都没有任何限制，只进行技术层面的适配。

（5）Pipeline运行。KPL平台为实现AI高效生产，同时满足开发者、普通用户的需求，开发实现了Pipeline功能。Pipeline是用户执行任务的“操作台”，每个Pipeline相互独立。用户在Pipeline内根据自身需求组合数据、算法、预训练模型、预处理器，调整相关参数，开始运行后任务会按照任务流图运行，并可查看任务运行中和历史任务详情。Pipeline创建后可反复使用。

（6）资源调度与管理。资源调度与任务管理：支持资源监控、模型训练等多种类型任务的统一调度管理，每种类型的任务均可根据资源需求实现动态调度，保证任务之间的资源共享与安全隔离。

名词解释

数据集

如果应用场景是对图像进行分析，那么数据集一般是由图片数据组成，每张图片可以称为一个样本，如果仅有图片，而没有标注信息，那么该数据集可以称为非标注数据集，如果有标注信息，那么称为标注数据集。标注就是通过人工等手段，对数据进行有目的性结构化描述，如何标注，标注哪些信息与具体的应用场景相关。比如目标是识别图片中是猫还是狗，那么标注就是对图片进行分类标注是猫还是狗。

如果数据集的用途是用于进行分类识别算法的训练，如上边说的分辨猫和狗，那么该数据集也可称为分类数据集。如果是用于检测算法的训练，如对人脸的位置进行定位，那么该数据集可称为检测数据集。诸如此类的还有语义分割数据集，实例分割数据集，人脸检测数据集，人脸识别数据集等。

数据集在被用于算法之前，常常需要对其进行处理，处理的目的一般有两个，一是为了适应算法对数据集结构化组织的要求，比如标注信息必须以XML进行存放，另一个是为了满足算法对数据读取性能的要求，在算法训练中，数据集越大，对数据的IO要求越高，如果以原始若干小文件的方式存储会拉低整个训练过程的速度。因此会对数据集进行重新序列化为一个或几个大数据文件（二进制），如TensorFlow中的TFRecords文件，进行IO加速。

模型

一般来说深度学习方法就是通过构造足够复杂的神经元节点及其节点连接，然后通过数据训练学习节点间连接的权重参数。最终训练的产出就是神经网络的结构和连接权重，一般我们会将其保存在一个或多个文件中，把这个文件称为模型文件，简称模型。模型就是算法训练过程最重要的产出物。

在训练一个全新的算法时，权重参数可以使用随机初始化。除此之外，还有一个更好的初始化方式，就是通过以往的模型进行初始化权重参数，但是前提是前后的网络结构具有相同部分，这样初始化会使得训练过程更快，而且可能还能取得精度上的提升。这种用途的模型，一般称作预训练模型，往往预训练模型是在一个超大数据集上训练出的高精度模型，能为加载它做参数初始化的算法训练提供很好的加速作用。

算法

算法可以将输入的数据变为可应用于我们期望场景的模型。从使用场景分，可以分为分类识别算法、检测算法、语义分割算法等等，每一大类下都有很多具体的算法。不同的算法几乎都在解决两个对立的问题，速度和精度，这是在生产应用中最关系的两个指标。从算法提供的功能上来看，有算法提供GPU的加速训练，甚至多机多卡的分布式训练，提供指标测试功能，提供API Server功能，提供将模型转换为适配某芯片模型的功能等等。在KPL中，我们将算法提供的功能分为了5种：训练，测试，批量推理，部署API，编译SDK。

Pipeline

Pipeline和下面要介绍的Notebook都是KPL提供的训练算法的两种使用方式。Pipeline通过拖拉拽节点（数据集，模型，算法）的方式构建任务。把模型，数据，算法等按期望的依赖方式进行组装，完成对数据进行预处理和训练算法等。

在下面几个算法例子小节中会专门介绍如何使用。

Notebook

如果希望即开即用的AI开发环境，弹性申请算力资源，编写AI算法代码，以交互式的方式调试代码，可以使用Notebook功能。Notebook是将JupyterLab进行了嵌入，并在环境中预置了不同的AI框架，如TensorFlow，PyTorch，MXNet等，方便您快速构建算法运行环境。

产品优势

（1）大规模数据训练的分布式架构。通过构建分布式训练架构，优化分布式IO，增加远程文件系统流式读取能力，处理海量数据，构建GPU多机多卡同步训练，实现在各种容器上高速运行大规模分布式训练。

（2）支持异构计算资源。平台支持多种异构资源管理，包括CPU、GPU、NPU、FPGA等，自研了融合管理平台，解决了多类型集群的管理难题，实现了混合调度，可将任务自动调度到异构计算资源中，最大程度的利用计算资源。

（3）全流程一体化的平台。构建训练-推理-应用的云化服务平台，提供覆盖全流程的开发套件，支持多框架、多硬件的组合，为用户提供高兼容性、高性能的部署能力，使用TensorRT/TVM等工具对深度神经网络进行特定于硬件设备和应用场景的模型压缩和加速，对Nvidia GPU、Intel CPU、寒武纪AI芯片等各种不同的计算设备平台进行优化，实现模型的快速部署。