

中科弘云深度学习计算服务平台 AI-Foundation产品简介

目录



01

人工智能平台介绍

02

产品简介

03

功能介绍

04

技术优势

05

应用场景及示例

传统人工智能面临的技术挑战

学习应用难

- Tensorflow、Caffe 等众多的计算框架以及 CNN、RNN 等复杂的网络模型，即便是资深工程师也需要花费大量的时间成本学习和应用。



管理调度难

- 主流计算框架采用 CPU+GPU/FPGA 的异构计算平台，其管理和调度融合了高性能计算、大数据和云计算等多领域技术，实现难度较大。



性能优化难

- 深度学习网络模型日趋复杂，通常包含数以万计的训练参数，如何进行超参数调优是提高训练推理效率的一个关键问题。



安装部署难

- 一套完整的计算环境包括操作系统、驱动程序、数学库、计算框架、网络模型、数据集等多个组成，有经验的工程师也通常需要1-2周才能完成



人工智能的新趋势--多技术融合



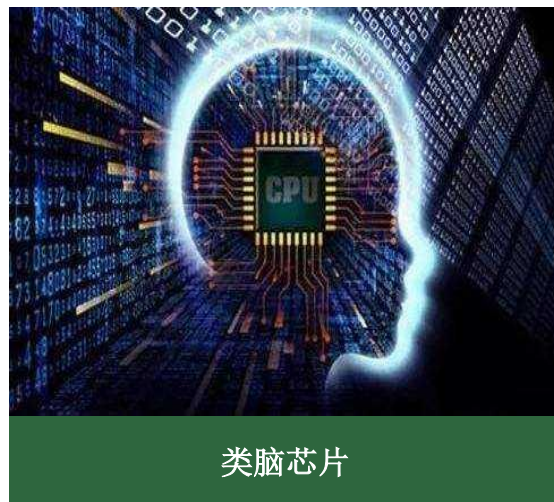
云计算



大数据



深度学习



类脑芯片

1

云计算

云计算使得成本低廉的大规模并行计算得以实现

2

大数据

大数据训练可以有效提高人工智能水平，为提高数据处理的效率和速度。

3

深度学习

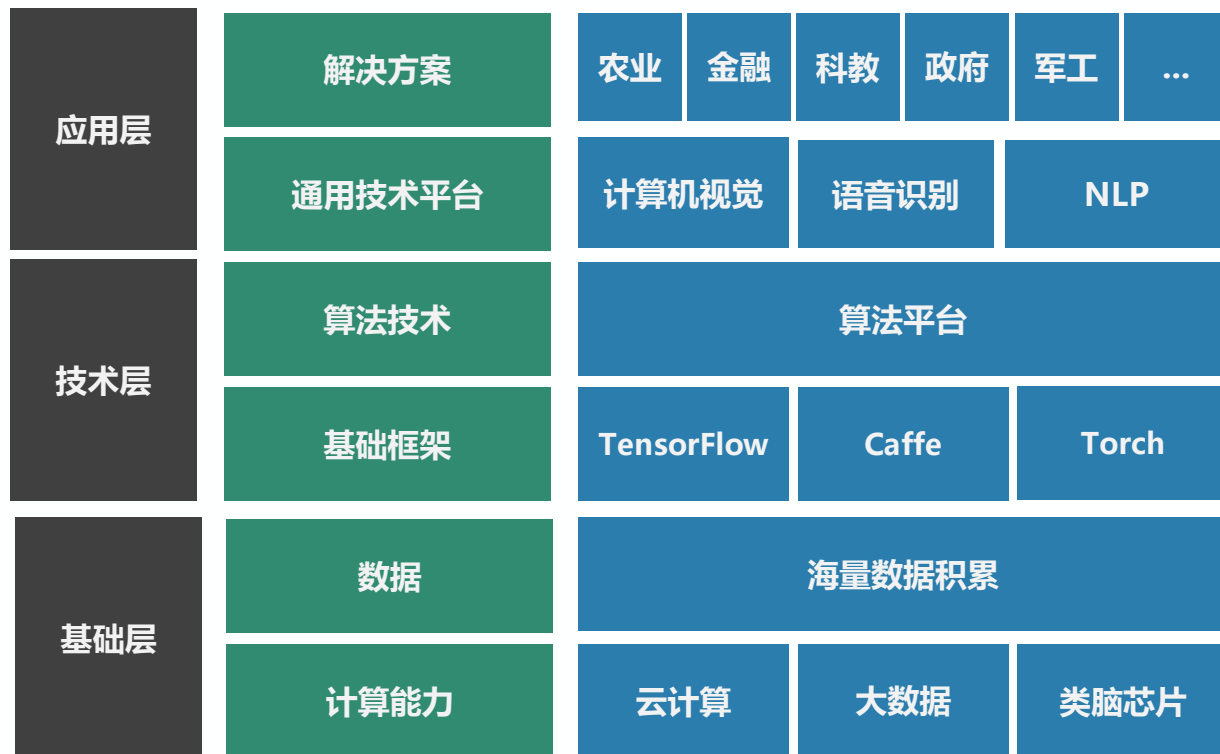
深度学习模拟人类大脑神经网络的工作原理，将人工智能带上了一个新的台阶

4

类脑芯片

人脑芯片也叫神经形态芯片，是从硬件方向对人脑物理结构的模拟

人工智能平台基础架构



◆ 基础层（按技术层级从上到下，下同）

计算能力层：大数据、云计算、GPU/FPGA等硬件加速、神经网络芯片等计算能力提供商。

数据层：身份信息、医疗、购物、交通出行等各行业、各场景的数据。

◆ 技术层

框架层：TensorFlow, Caffe, Theano, Torch, DMTK, 等框架或操作系统。

算法层：深度学习、增强学习等各种算法。

◆ 应用层

通用技术层：语音识别、图像识别、人脸识别、NLP等技术或中间件。

解决方案层：智能广告、智能诊断、自动写作、身份识别、智能投资顾问、智能助理、无人车、机器人等场景应用

目录

01

人工智能平台介绍

02

产品简介

03

功能介绍

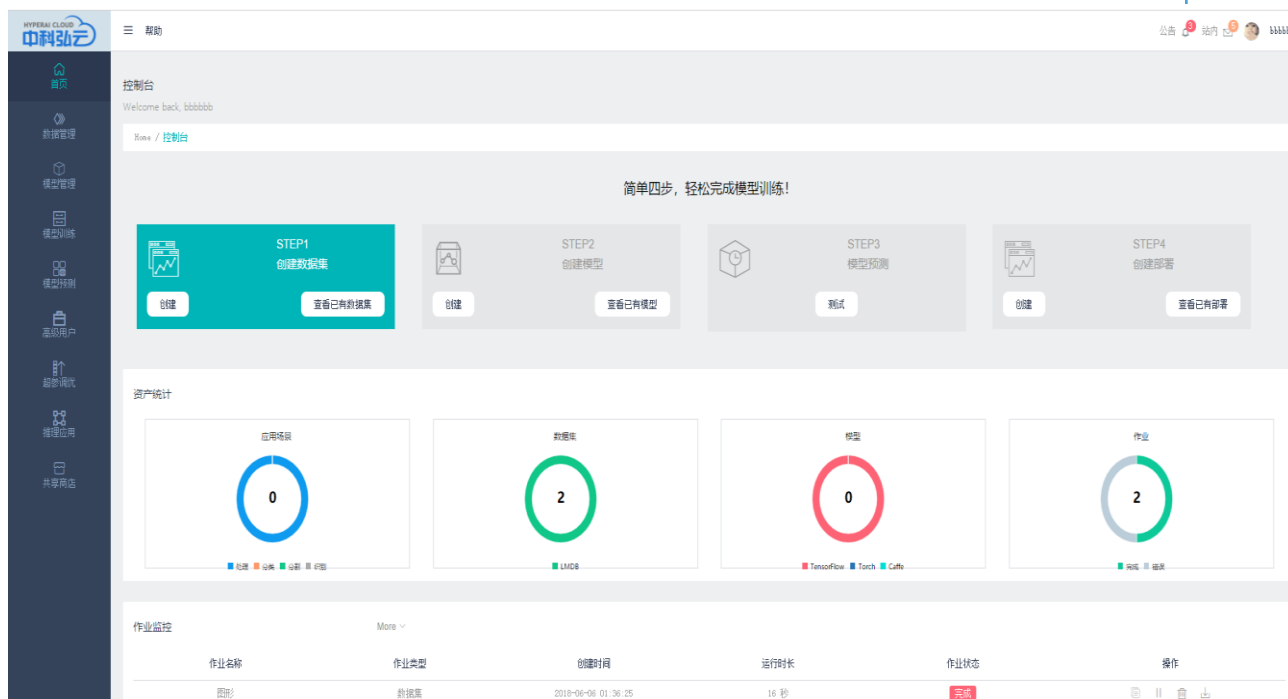
04

技术优势

05

应用场景及示例

AI-Foundation简介



AI-Foundation通过整合多种关键技术，提供从训练到推理的一站式人工智能云计算应用服务解决方案，能够帮助用户快速构建人工智能研发服务环境，大幅降低人工智能准入门槛，提升人工智能研发效率，用户无需编程也可获得强大的AI服务能力。

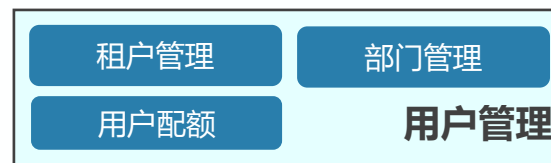
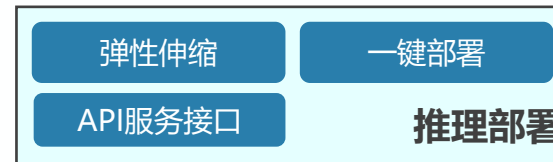
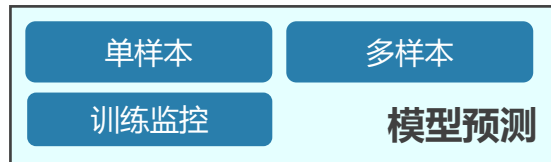
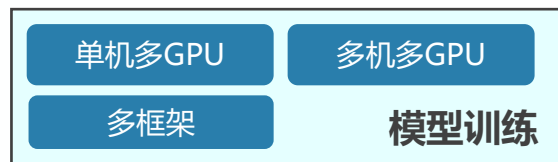
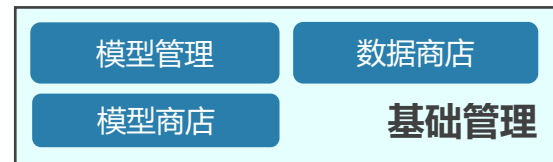
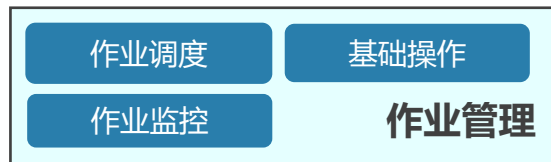
图像分类、目标检测、图像分割、医疗影像

数据集创建、模型训练、模型预测、推理部署

AI-Foundation--系统架构



AI-Foundation--功能模块



目录

01

人工智能平台介绍

02

产品简介



03

功能介绍

04

技术优势

05

应用场景及示例

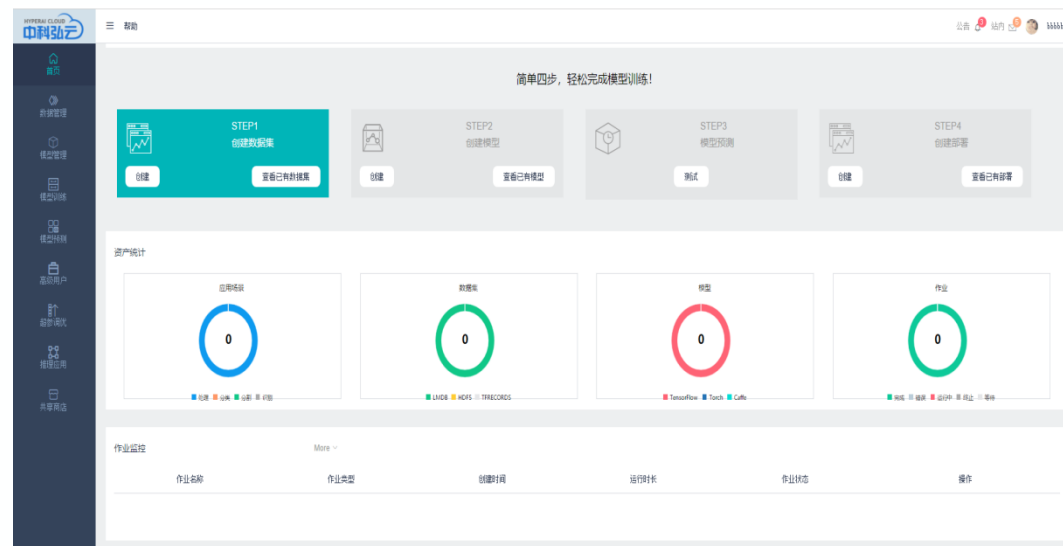
交互式全流程计算服务

简单易用的交互式界面

- 提供友好的WEB交互式服务界面，提供动态可视化训练作业监控图表和训练输出日志。

一站式全流程应用服务

- 通过让用户选择内置图像分类、物体识别、图像分割等应用场景、填写学习率、优化器等基本参数即可动态生成数据集创建、模型训练和推理预测等计算任务，用户全过程无需编写一行代码即可完成深度学习计算任务。



The screenshot displays the HYPERSAIL CLOUD web interface for creating a new dataset. The page title is "新建图像分类数据集" (New Image Classification Dataset). The form is divided into two main sections: "数据集属性" (Dataset Properties) and "图像属性" (Image Properties). The "数据集属性" section includes fields for "数据集名称" (Dataset Name), "描述" (Description), "上传数据集" (Upload Dataset) with a file selection button, "每批最小样本数" (Minimum samples per batch), "每批最大样本数" (Maximum samples per batch), "验证集比例" (Validation set ratio), and "测试集比例" (Test set ratio). There are also checkboxes for "验证集由全量数据集分割" (Validation set split from full dataset) and "测试集由全量数据集分割" (Test set split from full dataset). The "图像属性" section includes fields for "图像类型" (Image type), "图像尺寸(宽x高)" (Image size (width x height)), "尺寸变化" (Size change), "数据集类型" (Dataset type), "图像编码格式" (Image encoding format), "图像压缩" (Image compression), and "路径" (Path). The form has "取消" (Cancel) and "提交" (Submit) buttons at the bottom.

数据集管理

数据预处理

- 提供图像分类、目标检测、图像分割等典型场景的交互式数据集创建。支持上传样本数据、预测数据，提供图片归一化等数据预处理功能，

多种数据集格式

- 支持创建LMDB、HDF5、TFRecords等格式的数据集

数据集监控预览

- 支持数据集创建进度监控，支持数据集预览

HYPERAI CLOUD 中科弘云

公告 站内 用户

数据集属性

图像文件 文本文件

数据集名称:

描述:

上传数据集:

每类最小样本数:

每类最大样本数:

验证集比例:

测试集比例:

☐ 验证集目录是否单独隔离

☐ 测试集目录是否单独隔离

取消 提交

图像属性

图像类型: Color

图像尺寸 (宽x高): 256 x 256

尺寸变化: Squash

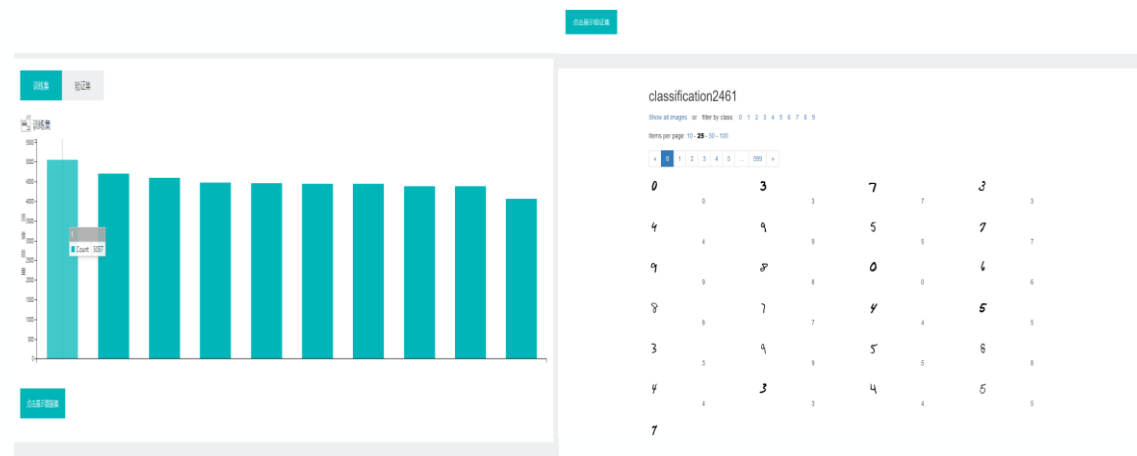
数据库属性

数据库类型: LMDB

图像编码格式: None

图像压缩: PNG (lossless)

组名: 请输入user-group



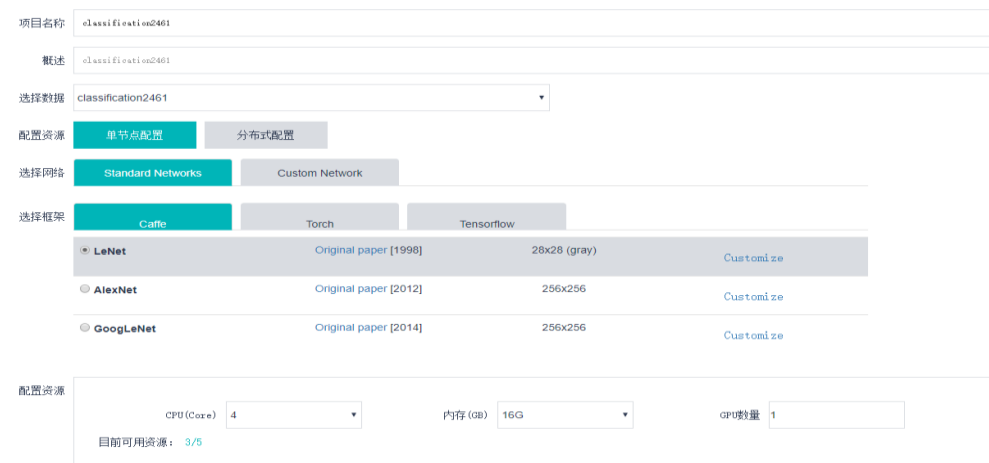
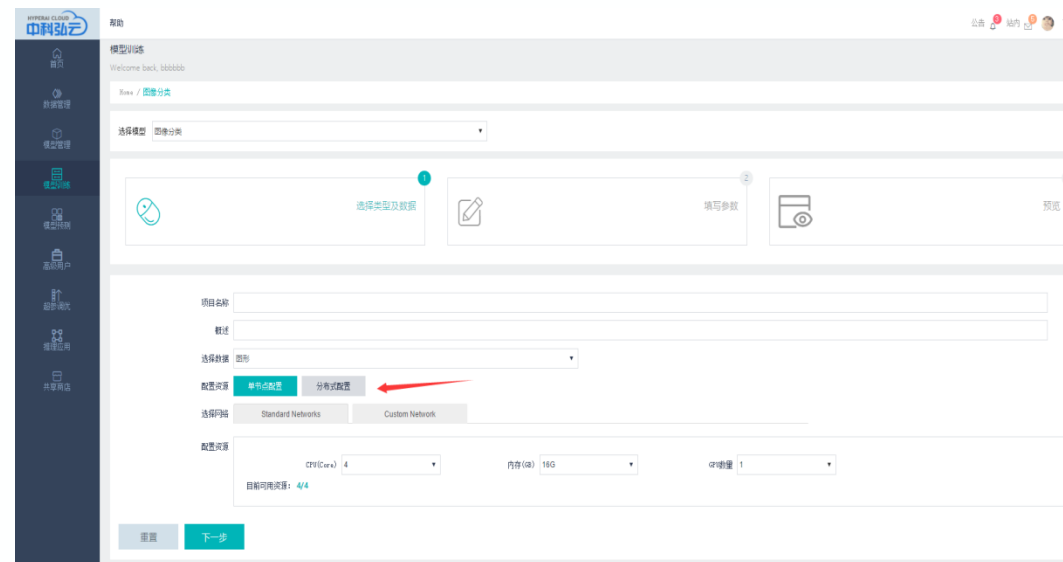
模型训练

支持模型并行训练

- 支持单机多GPU的并行训练任务
- 支持基于MPI的多机多GPU并行训练。

支持多种算法及框架

- 支持Tensorflow、Caffe、Torch、PyTorch、Caffe2等多种计算框架
- 支持GoogleNet, AlexNet等多种神经网络



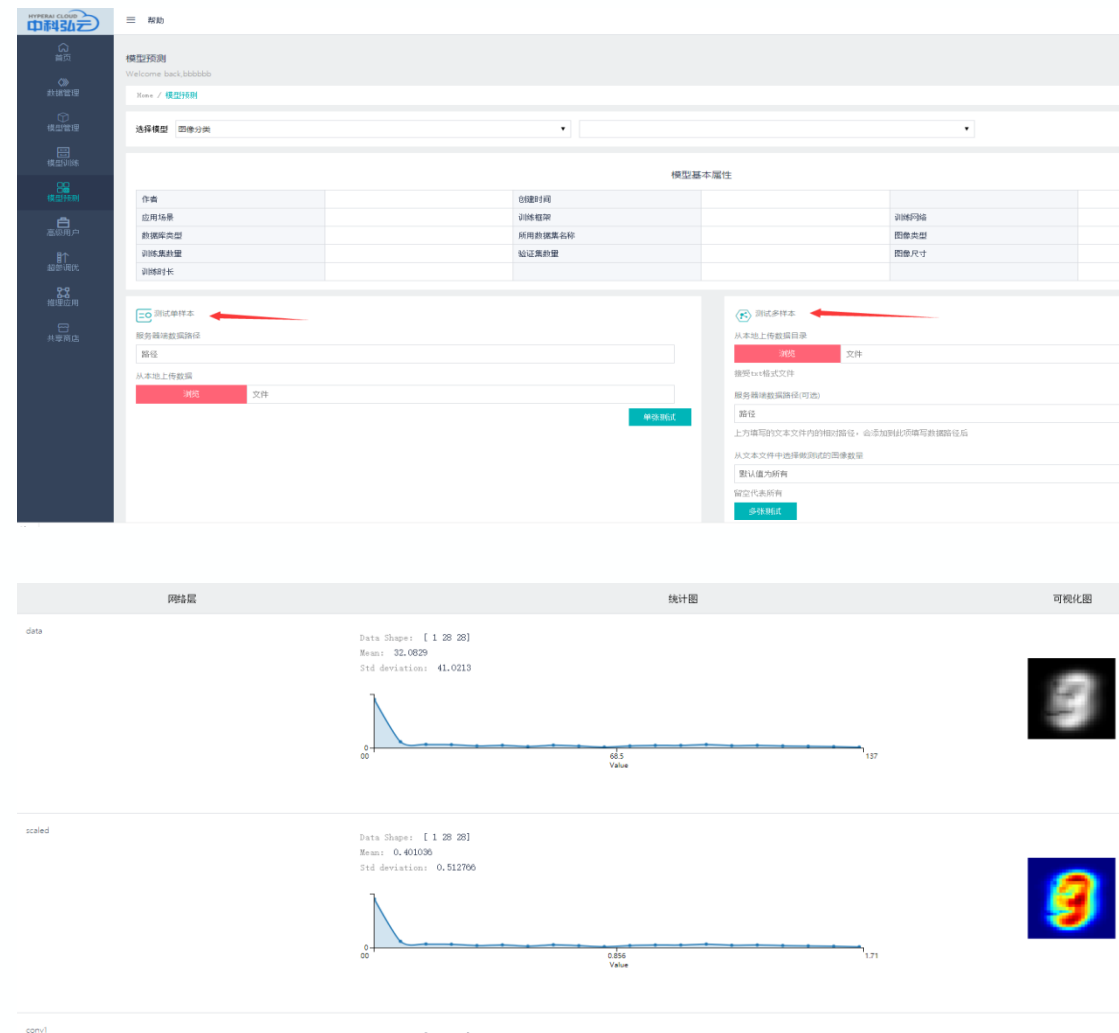
模型预测

多场景预测

- 通过选择场景和预训练模型，上传待测样本数据，实现单样本或多样本的预测任务

指标监控

- 提供单样本或多样本预测准确率指标输出，输出各层神经网络的参数图表，包含均值、标准差等。



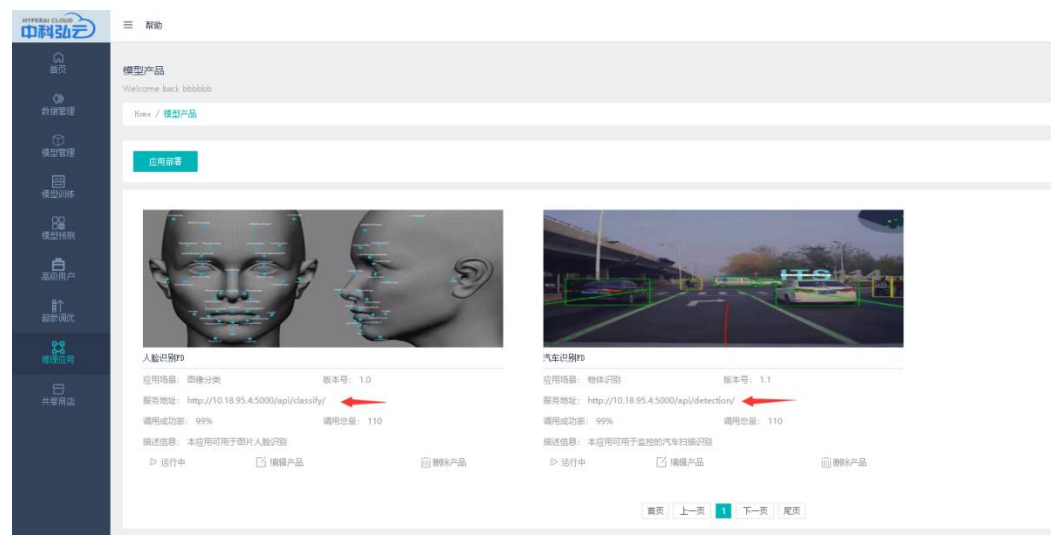
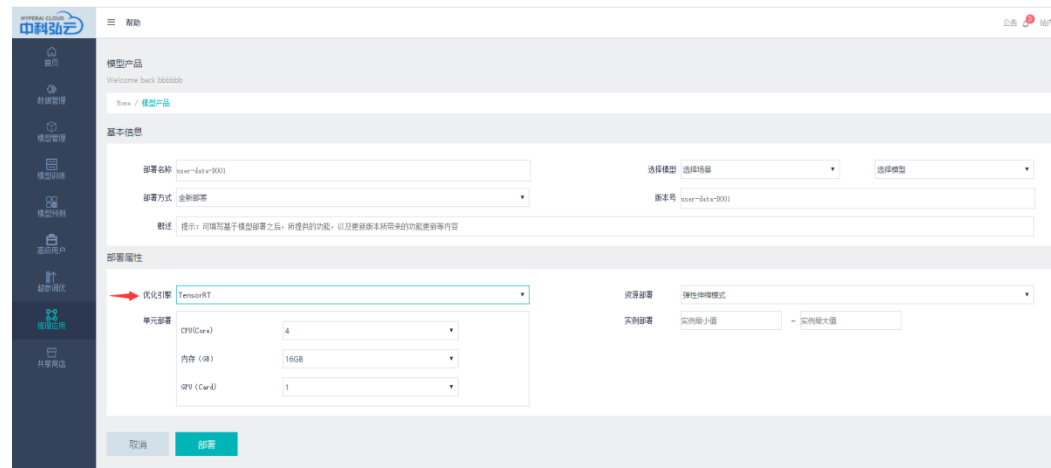
推理应用部署

一键部署

- 支持预训练模型的在线快速部署，采用TensorRT推理优化引擎，支持图像分类、物体识别场景下Caffe计算框架训练模型的直接一键部署。

API 接口

- 提供基于Restful API的模型调用服务接口。后续支持Tensorflow Severing推理引擎和Tensorflow模型部署。



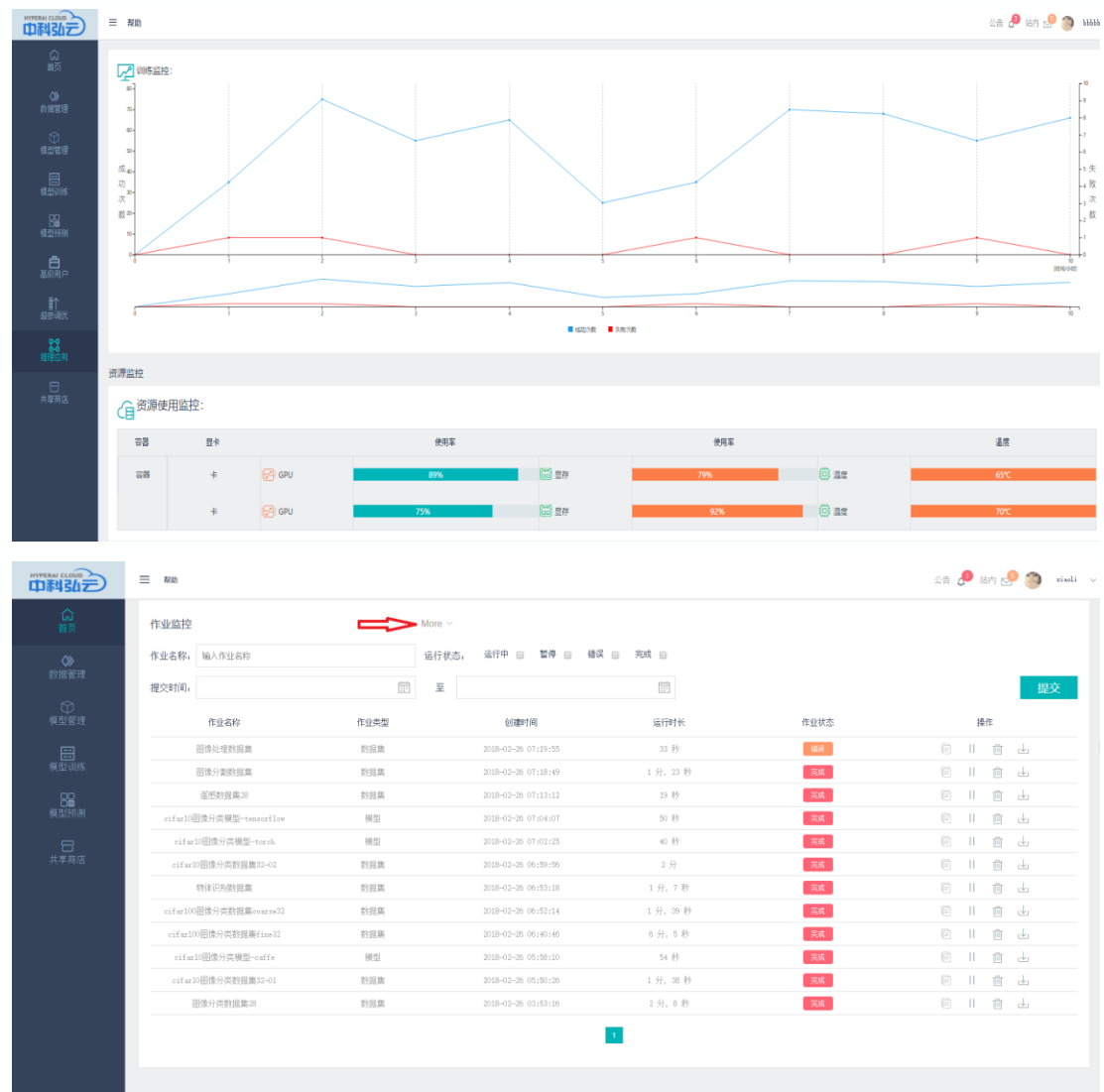
作业管理

作业监控

- 提供训练作业监控管理功能，包括作业运行百分比、作业运行状态（等待、运行、失败、结束等）、作业日志和图表输出、GPU资源监控等。

作业基础操作

- 提供作业快速克隆、作业删除、作业查询、作业终止等基本功能。



目录

01

人工智能平台介绍

02

产品简介

03

功能介绍



04

技术优势

05

应用场景及示例

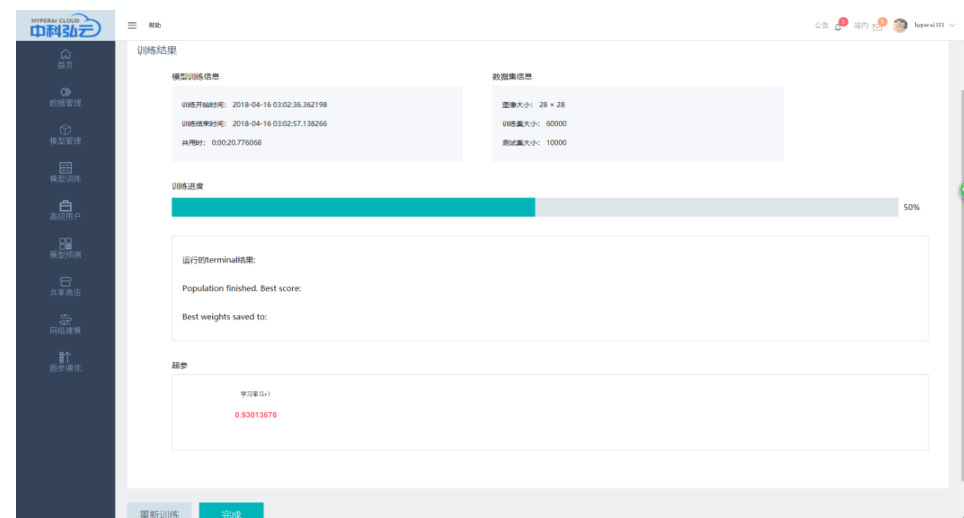
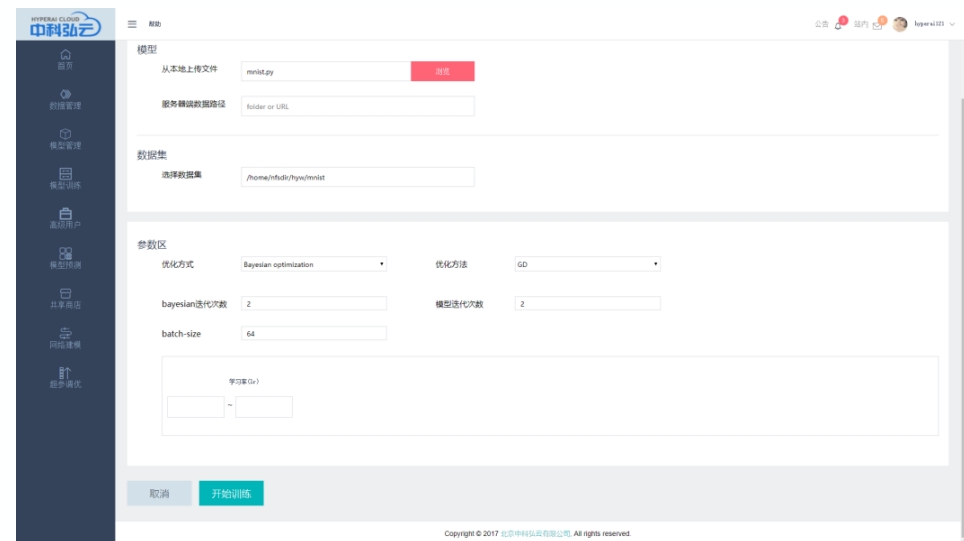
超参调优

功能描述

- 提供基于改进贝叶斯Bayes优化的超参数智能调优方法，支持GD、ADAM、Momentum等优化器，支持学习率、动量、指数衰减等超参数的智能调优。

优势和价值

- 从传统的调参过度到自动化的智能调参
- 能够快速的选择最佳的超参数集和模型



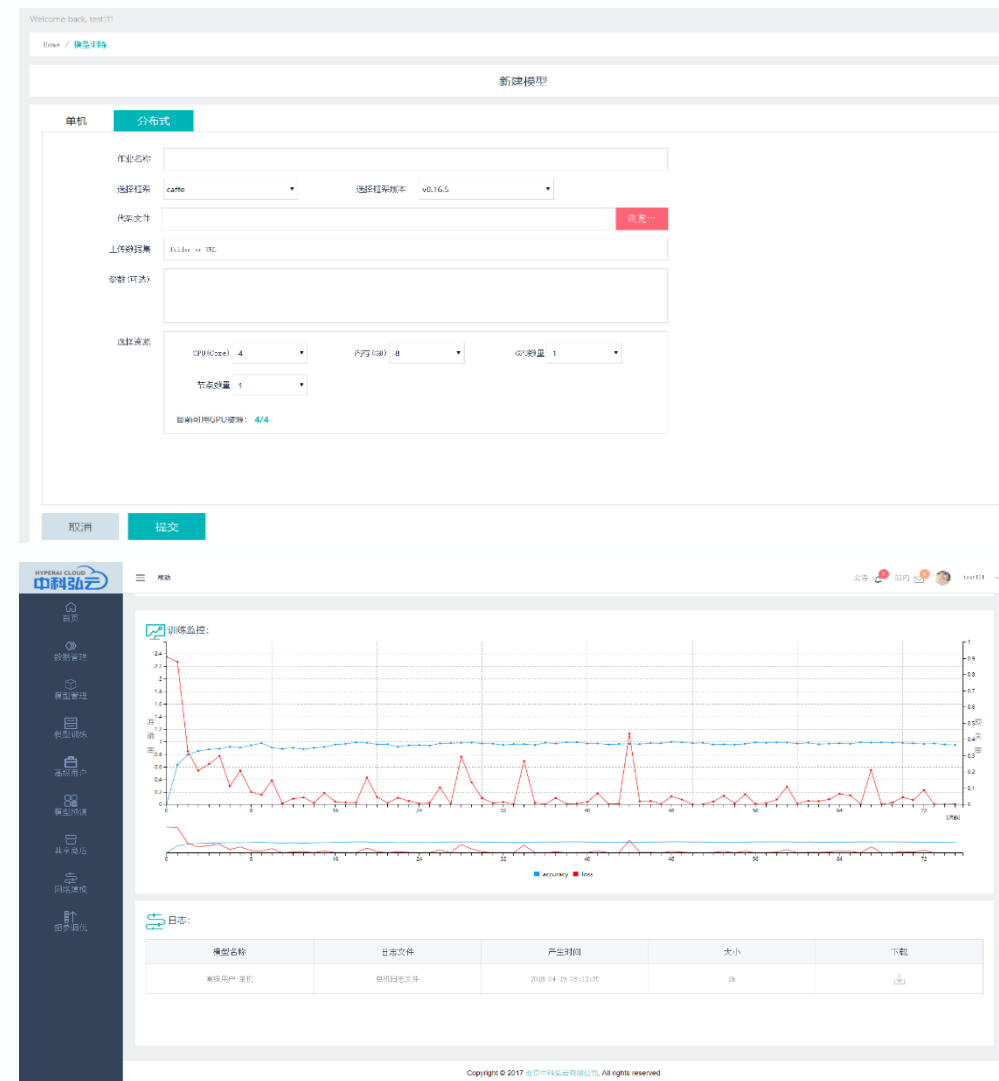
高级用户

功能描述

- 允许用户通过选择Tensorflow、Caffe等预制深度学习计算框架，并上传自定义的模型训练代码程序和训练数据集，即可启动单机或分布式并行训练任务。

优势和价值

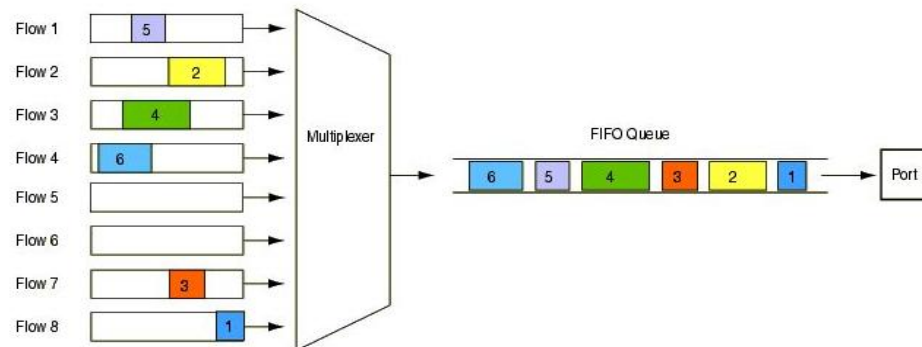
- 满足用户对于自定义模型训练代码的需求



作业调度

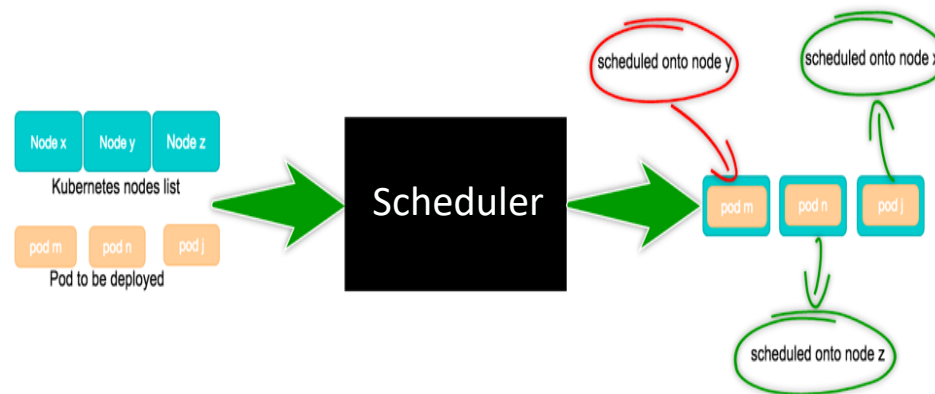
功能描述

- 提供动态资源调度管理功能，提供FIFO、作业优先级、资源配额、作业抢占等调度策略，支持自动选择性能最优的作业部署拓扑。



优势和价值

- 全面支持CPU、GPU等异构资源的混合调度
- 支持FIFO等多种作业调度策略



基于GPU的弹性伸缩

功能描述

- 支持推理部署的负载均衡模式，可根据GPU的负载变化情况动态调整计算资源，支持GPU使用率和显存的监控。

优势和价值

- 可满足用户对于突发性，大规模作业量的负载需求
- 可提高作业的容错能力和故障恢复能力

部署名称: test_01

部署方式: 全新部署

推理引擎: TensorRT

版本: 1.0

API部署: API部署图片上传

上传图片

选择模型: 图像分类

caffe-model-01

模型信息							
用户名	test_01	创建时间	2018-06-19 09:04:45	训练时长	200s	所用数据源名称	mnist-28
应用场景	classification	训练模型	caffe	训练网络	lenet	数据集类型	mnist
训练集数量	45002	验证集数量	14098				

部署模式: 弹性伸缩模式

实例数量: 2 ~ 4

GPU数量: 10 ~ 20

CPU (Cores): 4

内存 (GB): 16GB

GPU (Cards): 1



目录

01

人工智能平台介绍

02

产品简介

03

功能介绍

04

技术优势



05

应用场景及示例

应用场景



目标检测



图像分类



图像分割

图像分类--创建数据集

创建数据集

➤ 填写参数

名字: xxx

概述: xxx

上传数据集路径: xxxxxx

图像类型: 灰度图 (彩图)

图片大小 28x28 (256X256)

数据集属性

图像文件

文本文件

数据集名称: 图像分类01

概述: 图像分类

上传数据集: /home/nfsdir/hyperai_data/mnist/train

每类最小样本数: 2

每类最大样本数:

验证集比例: 25

测试集比例: 0

☐ 验证集与训练集是否隔离

图像属性

图像类型: Grayscale

图像尺寸 (宽X高): 28 X 28

尺寸变化: Squash

数据库属性

数据库类型: LMDB

图像编码格式: None

图像压缩: PNG (lossless)

组名: 请输入user-group

➤ 详情页

提交之后的详情页如图所示：

返回

图像分类01

发布者: testfor


应用场景: 图像分类

数据个数: 60000

图像分类

发布时间: 2018-05-06 08:19:10

数据库类型: lmdb



100%

训练集

验证集

图像分类--模型训练

模型训练

➤ 填写参数

进入填参界面，参数项如下：

名字： XXX

概述： XXX

数据集选择图像分类数据集

选择标准网络

项目名称 图像分类01caffe

概述 图像分类01caffe

脚本层

选择数据 图像分类01

配置资源 单节点配置 分布式配置

选择网络 Standard Networks Custom Network

选择框架 Caffe Torch Tensorflow

Network	Original paper	Resolution	Action
LeNet	Original paper [1998]	28x28 (gray)	Customize
AlexNet	Original paper [2012]	256x256	Customize
GoogLeNet	Original paper [2014]	256x256	Customize

配置资源

CPU (Core) 4 内存 (GB) 16G GPU数量 1

目前可用资源: 4/4

➤ 自定义网络

点击 standard networks 下网络对应的customize，跳转到 Custom Network页面

单节点配置 分布式配置

Standard Networks Custom Network

Caffe Torch Tensorflow

Visualize

```
1 # LeNet
2 name: "LeNet"
3 layer {
4   name: "train-data"
5   type: "Data"
6   top: "data"
7   top: "label"
8   data_param {
9     batch_size: 64
10  }
11  include { stage: "train" }
12 }
13 layer {
14   name: "val-data"
15   type: "Data"
16   top: "data"
17   top: "label"
18   data_param {
19     batch_size: 32
20  }
21  include { stage: "val" }
22 }
23 layer {
```

图像分类--模型训练

模型训练

➤ 其他参数设置

迭代次数自定义设置

其他参数默认即可

可对高级参数进行配置

通用参数

迭代次数间隔	5	模型自动保存间隔	30s	日志保存间隔	0s
快照间隔	1	验证间隔	1	跟踪间隔	0
随机种子	[none]	优化器类型	SGD (Stochastic Gradient Descent)		
基础学习率	0.01				
样本大小	[network defaults]				
样本处理批量					

学习率高级选项

模型参数

减平均	image	切片大小	none	参数3	12444
	image1	none			12444

重置 下一步

➤ 表单提交

进入训练监控页

训练监控页面如图：



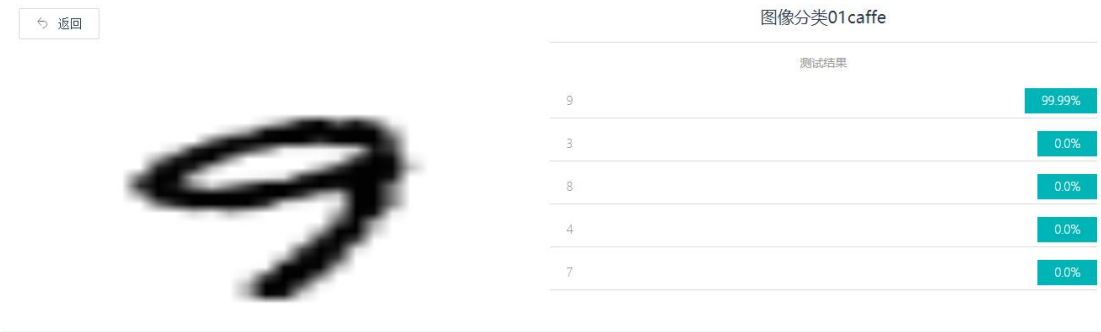
图像分类--模型预测

➤ 单张预测

上传图片的两种方式：

- ① 服务器路径上传：输入图片所在服务器端的绝对路径
- ② 本地上传：上传本地图片

点击单张预测，即可得到预测结果



➤ 多张预测

上传图片的两种方式：

- ① 本地上传多张图片
- ② 本地上传txt文档，服务器端路径填写所上传文档里图片所在服务器端的目录的绝对路径

点击多张预测，即可得到测试结果

图像分类01caffe

测试耗时: 测试数量: 3

序号	路径	测试结果	结果分布									
1	/home/nfsdir/testfor/infer/image1.png	9	9	99.99%	3	0.0%	8	0.0%	4	0.0%	7	0.0%
2	/home/nfsdir/testfor/infer/image2.png	2	2	99.98%	1	0.01%	7	0.0%	3	0.0%	0	0.0%
3	/home/nfsdir/testfor/infer/image3.png	0	0	100.0%	6	0.0%	2	0.0%	7	0.0%	5	0.0%

图像分类--模型部署

➤ 填写参数

部署名称: xxx (自定义)

概述: xxx (自定义)

部署方式: 选用全新部署

版本号: x.x (注意格式要求: 版本号为数字, 如1.0)

推理引擎: TensorRT (支持TensorRT和TF serving)

模型上传: 本例选择模型上传下的服务器路径

部署模式: 本例选用弹性部署

模型信息			
用户名	test_hy	创建时间	2018-06-22 08:10:44
应用类型	classification	训练框架	caffe
训练集数量	45002	验证集数量	14998
训练时长	57s	所用数据集名称	mnist
训练网络	lenet	数据类型	imdb

资源属性	
部署模式	弹性伸缩模式
实例数量	2 ~ 4
CPU (Core)	4
内存 (GB)	16GB
GPU (Card)	1

➤ API调用 (预测)

在已完成部署的API监控 (详情) 页面, 点击服务地址栏的链接可直接进入API 的调用 (预测) 页面。

点击上传图片, 选择需要预测的图片, 即可进行预测。

结果:

```
{
  "confidence": 0.9845, "label": "2"
}, {
  "confidence": 0.0134, "label": "7"
}, {
  "confidence": 0.002, "label": "1"
}
```

成功次数: 2
失败次数: 0



THANK YOU!