# 产品介绍

[Mo 模型管理系统](https://modelshop.atom.cg.supremind.info)，是Ti深度学习系统的补充。

一方面：Ti系统训练完成的模型，需要经过专业的测试评估才能交付使用。此处为了更全面客观的对模型进行评估，会由测试工程师的另外根据需求场景进行测试用例集合的收集并对模型进行评估。而 Mo 模型管理系统就提供了方便测试工程师维护测试数据集并对模型推理评估的功能；

另一方面：Ti系统提供python模型的训练，然而python模型的运行效率并不是最好的。因此为了提升模型的运行性能，闪马人工智能研究院推出了自研的Tron模型架构，将python模型的性能提升了50%以上。而Mo模型管理系统，正是提供了模型转换封装的能力。

### 产品功能

1.支持Python模型转换Tron模型功能；  
2.支持Tron模型转换盒子模型（适用于华为Ascend硬件平台的Tron模型）功能；  
3.Tron模型量化；  
4.Tron模型评估；  
5.Tron模型验证/测试功能；  
6.Tron模型发布管理；

### 产品特色

1.支持多种模型处理工具  
- 支持对不同类型的模型进行转换，转换后适用于Nvidia GPU, Huawei Ascend多种平台；  
- 通过模型量化对模型进行优化：减少内存和存储占用，降低功耗，提升计算速度；

2.高效的模型管理方式： - 系统内用户可以上传相同模型的不同版本； - 用户在数据集管理模块通过不同版本模型对相同的数据集的推理精度；

# 二、名词概念

为了让入门者更好的使用Mo模型管理系统，或者说更好的使用Mo进行模型转换、推理，必须了解相关的基础概念。用户手册中包含有关模型和训练等相关概念，为了让用户更好的理解用户手册的内容，本章会着重对这些名词进行解释。

后文中涉及的模型转换，模型量化，模型推理的概念，需要用户在使用训练功能之前对这些内容做好了解的基础知识；

# 有关模型和训练的名词

### 模型结构（网络结构）

主干（骨干）网络（backbone），一般是指提取特征的网络，其作用就是提取图片中的信息，供后面的网络使用。此外，被称作backbone的网络通常都是已在大型数据集(例如ImageNet、COCO等)上完成了预训练，拥有预训练参数的卷积神经网络。

经常使用的主干网络有ResNet、VGG等，这些网络已经被充分证明在分类等问题上具有很强的特征提取能力。在使用这些网络作为backbone的时候，都是直接加载官方已经训练好的模型参数，后面接着我们自己的网络。通常会同时训练网络的这两个部分，这是因为加载的backbone模型已经具有提取特征的能力了，而在我们的训练过程中，只需要对其进行微调，使得其更适合于我们自己的任务。

### epoch

使用训练集的全部数据进行一次完整的训练，简称“一代训练”。一个Epoch意味着训练数据集中的每个样本都有机会更新内部模型参数。Epoch由一个或多个Batch组成。

image

image

在训练的过程中，随着epoch数量的增加，神经网络中权重更新迭代的次数增多，曲线从最开始的不拟合状态，慢慢进入优化拟合状态，最终进入过拟合。

### batch

所有训练样本在神经网络中都进行了一次正向传播和一次反向传播。也就是1个epoch等于使用训练集中的全部样本训练一次。

### batch size

每批样本的大小，即每次训练在训练集中取batchsize个样本进行训练。算法工程师通过训练模型结构的网络大小和resize后图片大小，再根据训练所用的计算资源内存的大小进行判断，确定的batch size；盲目增大batch size，虽然内存利用率提高了，但是显存容量可能撑不住，报错CUDA OOM。

### Iteration

1个iteration等于使用batch size个样本训练一次。一个迭代=一个正向通过+一个反向通过。训练一个Batch就是一次Iteration

举个例子：

假设训练集有1000个样本，batchsize=10，那么训练完整个样本集需要：

* 100次iteration
* 100个Batch
* 1次epoch

# 模型

### 什么是PTH模型？

在Ti深度学习系统中基于Pytorch框架训练的模型，训练平台输出的通常是权重的优化，几乎没有训练会使用模型的网络结构变化的策略，例如剪枝等；等待训练执行到设置好的epoch会输出的模型的评估和存储下来的模型文件，存储下来的模型就是PTH模型；

我们这里说的PTH模型只保存了模型的各个操作符和具体的参数值，没有保存拓扑结构；

### 什么是Tron模型？

Tron模型是根据PTH模型通过压缩参数和一些运算方式得到的，保存了原本的拓扑结构，具体的模型转换方式请参考下文的“模型转换模版”；

有两个硬件平台需要运行tron模型：  
\* 运行在英伟达GPU硬件上的Tron模型：由PTH模型直接转换得到的Tron模型可以直接创建模型推理任务，运行的硬件平台在英伟达GPU上；

* 运行在华为Ascend盒子上的Tron模型：由于运行在华为Ascend硬件平台上，所以输入以不同色彩空间存储的图片，所以我们也要对模型所输入的图片的色彩空间进行转换；由于模型网络不同也需要对，mean\_value、scale\_value这两个参数模型读入数据进行归一化操作的两个参数也有所不同，因此需要重新设置；

# 量化

### 什么是量化？

模型量化就是将浮点存储（运算）转换为整型存储（运算）的一种模型压缩技术。简单直白点讲，即原来表示一个权重需要使用float32表示，量化后只需要使用int8来表示，同时我们期望转换后的模型准确率与转化前相近。

仅仅这一个操作，我们就可以获得接近4倍的网络加速！

### 模型量化有什么好处？

* 减少内存和存储占用： 量化对模型的『瘦身』效果可谓立杆见影，它所带来的好处是两方面的：其一，减少memory footprint。我们知道很多时候推理性能的瓶颈不在于计算而在于访存，在这种情况下提高计算密度会对耗时有显著的优化效果；其二，节省存储空间，减少应用的size，同时便于软件的升级更新。
* 提升计算速度： 很多处理器整数计算指令都要比相应的浮点计算高效。以CPU为例，浮点运算指令的latency平均会长于对应的整数运算指令。

### 量化的对象是什么？

一般来说，会对模型的哪些数据进行量化。主要有以下三个。实际中可能是量化其中的多个甚至全部。

* weight（权重）： weight的量化是最常规也是最常见的。量化weight可达到减少模型大小内存和占用空间。
* activation（激活函数输出）： 实际中activation往往是占内存使用的大头，因此量化activation不仅可以大大减少内存占用。更重要的是，结合weight的量化可以充分利用整数计算获得性能提升。

### 量化的位数

大部分的科学运算都是采用浮点型进行计算，常见的是32位浮点型和64位浮点型，即float32和double64。

对于量化的位数，可以有很多种选择。大体可分为几类：

* 16位
* 8位 最常见也相对成熟。各种主流框架和硬件都支持。
* 8位以下目前而言学界相对玩得多些，工业界有少量支持，但还没有太成熟。8位以下主要是4，2和1位（因为位数为2的幂次性能会更好，也更容易实现）。如果精度低至1位，也就是二值化，那可以用位运算进行计算。这对处理器而言是很友好的。

# 模型转换

为什么要模型转换？转换些啥？

## 英伟达GPU模型转换

PTH模型是没有办法直接用来跑推理测试任务的，因为PTH模型只包含了模型的操作符和参数值没有模型的结构；而运行推理测试的模型需要封装模型结构；

因此，我们将PTH转换成了Tron模型，里面封装了模型结构，操作符和执行参数，用于跑推理测试任务；转换好的Tron模型就可以在英伟达GPU上跑推理测试任务

由于运行的平台不同，对模型的参数要求也不同：

比如在英伟达GPU和华为Ascend盒子上，由于输入图片的标准不同，所以在转换成不同平台运行的模型时，输入图片的标准化的参数参数和配置有所不同，因此在模型转换的时候需要添加数据格式转换的操作；

## 华为Ascend盒子模型转换

色彩空间的转换 由于模型推理需要输入以不同色彩空间存储的图片，所以我们也要对模型所输入的图片的色彩空间进行转换，目前支持下列的的图像通道转换：  
\* rgb->bgr  
\* bgr->rgb  
\* nv12->bgr  
\* nv12->rgb

### 归一化设置

mean\_value、scale\_value这两个参数是对模型读入数据进行归一化操作的两个参数配置，模型按照设定，对输入数据的预处理计算的系数；

### 日志的输出内容配置

在转换成华为Ascend盒子模型的过程中，转换脚本运行时，输出的日志哪些需要展示出来在portal页面上提供给用户查看；  
目前有debug，info，error，warning四种日志的输出，支持多选，不选的话默认配置为None；

# 模型推理测试

### 模型推理的输入

在创建模型推理测试任务时输入的参数：  
1. 需要选好推理测试任务运行的平台，是英伟达GPU还是华为Ascend盒子。  
2. 并选好运行的模型和数据集，其中快速验证这一选项是用本地上传的图片跑推理测试任务，不会与数据集产生关联，这是用来验证模型能否用这类数据集跑通的测试。  
3. 设置好推理测试任务每次dataloader取数据到内存中进行推理的batch size。  
4. 模型循环的次数。  
5. 评估IOU阈值以及推理测试任务使用的模型的精度。

image

image

### 模型推理测试输出

模型推理测试执行之后会输出任务的运行详情，包括任务的评估IOU阈值，batch size和推理运行完成的总耗时等任务的详细信息；  
以及对目标推理的推理图、不同标签的推理精度统计以及整体的精度统计；  
image

### 推理参数

这些推理结果可以用来测试模型的精度，也可以以数据集的维度去查看和对比不同版本的模型，对同一数据集进行推理，推理的精度差别，任务总时长，预处理时长，推理时长和后处理的时长，用来对比不同版本模型性能上的区别。  
image

更多产品细节请见第三章使用指南

# 三、快速入门

算法用户需要测试不同硬件平台的模型，因此在使用流程上会有所不同，根据不同的用户使用需求，我们分开介绍这两种硬件平台的模型测试流程；

## 英伟达GPU硬件平台

PTH模型是没有办法直接用来跑推理测试任务的，因为PTH模型只包含了模型的操作符和参数值没有模型的结构；而运行推理测试的模型需要封装模型结构； 因此，我们将PTH转换成了Tron模型，里面封装了模型结构，操作符和执行参数，用于跑推理测试任务；转换好的Tron模型就可以在英伟达GPU上跑推理测试任务；  
具体操作流程：[英伟达GPU硬件平台](/model-shop/getting_started/NVIDIA)。

## 华为Ascend硬件平台

由于运行的平台不同，对模型的参数要求也不同， 在英伟达GPU和华为Ascend盒子上，由于输入图片的标准不同，所以在转换成不同平台运行的模型时，输入图片的标准化的参数参数和配置有所不同，因此在模型转换的时候需要添加数据格式转换的操作； \* 色彩空间的转换：由于模型推理需要输入以不同色彩空间存储的图片，所以我们也要对模型所输入的图片的色彩空间进行转换，目前支持rgb->bgr，bgr->rgb，nv12->bgr，nv12->rgb 支持模型在模型读取图片的时候模型的图像通道BGR转成RGB； \* mean\_value、scale\_value这两个参数是对模型读入数据进行归一化操作的两个参数配置： 模型按照设定，对输入数据的预处理计算的系数； \* 日志的输出内容配置： 在转换成华为Ascend盒子模型的过程中，转换脚本运行时，输出的日志哪些需要展示出来在portal页面上提供给用户查看； 目前有debug，info，error，warning四种日志的输出，支持多选，不选的话默认配置为None；  
具体操作流程：[华为Ascend硬件平台](/model-shop/getting_started/ASCEND)。

# 英伟达GPU硬件平台

算法用户在训练平台会基于pytorch、tensorrt等深度学习框架进行模型优化和迭代，输出的PTH模型文件，只包含了模型的操作符和参数值没有模型的结构； 为了让模型可以成功在英伟达GPU硬件平台上跑推理测试任务： \* ModelShop内置了模型转换功能，封装、转换成可以在英伟达GPU、华为Ascend硬件平台上运行的Tron模型； \* 由于硬件的算力和内存的限制，让一些较大的模型可以在ModelShop运行推理测试任务，也需要对这些模型进行压缩，所以系统也具备对模型进行量化的功能； 下面是算法用户在英伟达GPU硬件平台运行推理测试任务整体的操作流程图； image 使用流程概述：

1.[模型准备](/model-shop/instruction/model_management)：用户在进行模型测试前，需要上传测试的模型到 Mo 上，可以直接上传转换好的英伟达Tron模型，也可以上传PTH模型转换后使用；如果模型较大，硬件运行起来较为吃力，或者需要对模型网络进行加速，可以使用[模型量化](/model-shop/instruction/quantization)功能，完成量化后，可以使用int8的模型运行推理测试任务；

2.[模型转换](/model-shop/instruction/model_transfer)：如果用户上传的是PTH模型需要在 Mo 中转换，完成模型转换的到转换好的英伟达Tron模型才可以使用这个tron模型创建英伟达GPU硬件平台的模型推理测试任务；

3.[数据集管理](/model-shop/instruction/dataset)：在 Mo 上进行模型推理测试，用户需要准备用于推理的数据集用户可以在DataOcean上选择LabelX上已经归档的标注任务、标注包，创建数据集，DataOcean可以对生成的数据集进行格式转换，转成 Mo 中可以用于推理的数据集格式；

4.[创建推理测试任务](/model-shop/instruction/infer)：创建模型推理测试任务，用户可以根据不同的测试要求对任务进行配置，例如：batch size（一次训练所选取的样本数）、循环次数（一般用于单张图对模型进行测速），评估IOU阈值（评估任务在数据集中检测相应物体准确度的一个标准）等参数；

5.查看推理结果：推理测试任务运行完成后，可以查看推理测试任务的推理精度和推理任务的不同指标；

# 华为Ascend硬件平台

算法用户在训练平台会基于pytorch、tensorrt等深度学习框架进行模型优化和迭代，输出的PTH模型文件，只包含了模型的操作符和参数值没有模型的结构； 为了让模型可以成功在英伟达GPU硬件平台上跑推理测试任务： \* ModelShop内置了模型转换功能，封装、转换成可以在英伟达GPU、华为Ascend硬件平台上运行的Tron模型； \* 由于硬件的算力和内存的限制，让一些较大的模型可以在ModelShop运行推理测试任务，也需要对这些模型进行压缩，所以系统也具备对模型进行量化的功能； 下面是算法用户在英伟达GPU硬件平台运行推理测试任务整体的操作流程图； image

使用流程概述：

1.[模型准备](/model-shop/instruction/model_management)：用户在进行模型测试前，需要上传测试的模型到 Mo 的模型管理模块中，可以上传在英伟达GPU运行的TRON模型，使用模型转换功能，转换成可以在华为Ascend硬件平台上运行的Tron模型；也可以直接上传线下通过脚本转换好的华为Ascend平台的Tron模型；

2.[模型转换](/model-shop/instruction/model_transfer)：如果用户上传的是上传在英伟达GPU运行的TRON模型，使用模型转换功能，创建模型转换任务，生成成可以在华为Ascend硬件平台上运行的Tron模型；

3.[数据集管理](/model-shop/instruction/dataset)：在 Mo 上进行模型推理测试，用户需要准备用于推理的数据集用户可以在DataOcean上选择LabelX上已经归档的标注任务、标注包，创建数据集，DataOcean可以对生成的数据集进行格式转换，转成ModelServer中可以用于推理的数据集格式；

4.[创建推理测试任务](/model-shop/instruction/infer)：创建模型推理测试任务，用户可以根据不同的测试要求对任务进行配置，例如：batch size（一次训练所选取的样本数）、循环次数（一般用于单张图对模型进行测速），评估IOU阈值（评估任务在数据集中检测相应物体准确度的一个标准）等参数；

5.查看推理结果：推理测试任务运行完成后，可以查看推理测试任务的推理精度和推理任务的不同指标；

# 四、使用详解

本章节会对系统中每个功能点进行展开详细的说明，如果您对某一功能点的使用方法不太清楚，可以在本章中找到对应模块的使用详解，对照使用步骤的详细说明进行操作； - 使用前，需要完成账号注册并登录，详细操作指导可参考； - 如果您需要在英伟达GPU上运行推理测试任务，模型的全部流转及操作过程；  
详细操作指导可参考[NVIDIA平台用户指南](/model-shop/getting_started/NVIDIA)； - 果您需要在华为Ascend上运行推理测试任务，模型的全部流转及操作过程；  
详细操作指导可参考[ASCEND平台用户指南](/model-shop/getting_started/ASCEND)；

# 1.模型管理

用户在进行模型测试前，需要上传需要测试的模型到 Mo 上，Mo 支持上传3种类型的算法模型，分别是：PTH、英伟达GPU Tron模型和华为Ascend平台模型；

训练平台生成的PTH模型只有模型的权重和操作符，没有模型结构；为了能运行模型推理测试任务，算法人员需要上传PTH到模型管理模块，完成模型结构的封装和参数的压缩后，才能使用转换好的Tron模型执行推理测试任务；因此上传PTH和英伟达GPU模型后需要将这两类模型进行模型转换，按照与算法约定的转换规则转换成Tron模型文件；  
详细步骤见下文：[模型转换](/model-shop/instruction/model_transfer)

# 1.1 PTH模型管理

## 1.1.1 上传PTH模型

为了进行模型推理，我们需要先把pth模型上传到 Mo 中，并转换成tron model，并通过转换好的tron model和从DataOcean中转换到 Mo 的数据集才可以启动模型的推理任务

步骤1:进入模型管理-PTH模型-新建界面，在左侧栏目中进入“模型管理”模块，进入PTH模型这一栏，在“我的”这一tab页面中点击“新建”，上传需要使用的PTH模型，

image

image

步骤2:模型文件上传

在上传PTH模型时，需要输入模型名称（必填项），上传文件类型为pth的模型文件（必填项），以及模型文件的说明（选填项） image

步骤3:查看上传模型

上传成功后即可在PTH模型的“我的”和“全部”tab页面中查看到上传的模型和相关信息：模型ID、模型名称、训练ID（目前暂不支持，训练ID来自于训练归档到DataOcean的TP训练任务，便于对模型的跟踪），说明，创建模型日期（上传该模型的日期），以及下载该模型的功能。由于字段较多页面锁定了操作栏，其他字段可以左右滚动查看。

image

image

## 1.1.2 编辑PTH模型

如果需要修改上传的PTH模型的信息，可以在模型列表页面，找到需要修改的模型点击右侧操作栏中的“编辑”修改模型名称和模型的说明；

image

image

## 1.1.3 下载PTH模型

如果用户需要下载PTH模型到本地，点击操作栏中的“下载”按键，浏览器会自动下载这个模型到本地；

## 1.1.4 删除PTH模型

如果PTH模型需要删除，可以点击对应模型操作栏中的“删除”按键对模型进行删除；删除模型需要二次确认；

image

image

# 1.2 NVIDIA G模型管理

在英伟达硬件平台运行的Tron模型可以直接通过上传的方式新建，也可以通过已经上传的PTH模型进行转换得来，如何用PTH转换Tron model会在下面的章节3中详细介绍。 ## 1.2.1 上传NVIDIA 模型 步骤1:进入模型管理——Tron模型这一栏目，并点击“新建”按钮 image

步骤2:上传英伟达模型

需要上传Tron model的相关信息，系统会根据下面填写的信息按照一定规则组合在顶部“模型文件名称”这一栏自动生成Tron model的模型文件名称。

* Repo：模型的类别（必选项，展示默认值）
* 模型名称：用户自行命名的模型名称（必填项，对输入内容无限制）
* method（模型的网络格式）：模型的推理后处理方法（必选项，展示默认值）
* 版本：模型的版本信息（必填项，对输入内容无限制）
* 模型文件：在这里上传的Tron model文件。（必填项，不会对上传文件格式进行校验，请上传Tron model文件）
* 选择模版：这个Tron model是由哪个模型转化模版转化得来的（非必选项，模型转化模版是在模型转化这一模块，在模型转化这一页面上传得来，如未找到使用的转化模版，请上传转化模版后再上传tron model）
* 说明：对这个模版的描述，会在Tron model列表中展示（非必填项，对输入内容无限制）

image

image

步骤3:查看Tron model上传或转化的Tron model都会在这里展示

在模型列表中可以查看模型的“描述”，“下载”，“量化”和“删除”等操作。下文会详细介绍这些操作的作用和用法；

image

image

## 1.2.2 查看模型描述

“描述”为展示对应模型的json文件，采用语法控件，目前后端接口暂不支持；

## 1.2.3 下载模型

果用户需要下载PTH模型到本地，点击操作栏中的“下载”按键，浏览器会自动下载这个模型到本地；

## 1.2.4 量化模型

目前系统只支持对英伟达GPU平台上的Tron模型进行模型量化，且该模型需是从PTH模型转换成Tron模型才能进行模型量化；直接上传的英伟达GPU平台上的Tron模型无法进行量化，“量化”操作处于disable的状态；下文会详细介绍如何对PTH模型转换成Tron模型进行模型量化；

## 1.2.5 删除模型

模型需要删除，可以点击对应模型操作栏中的“删除”按键对模型进行删除；删除模型需要用户二次确认，防止误操作的发生；

image

image

## 1.3 华为Ascend模型管理

## 1.3.1 上传华为Ascend模型

在华为Ascend盒子硬件平台上运行的模型可以通过“新建”的方式直接上传；

步骤1:点击左上角的“新建”按钮通过直接上传的方式创建一个可以在华为Ascend盒子上运行的模型； image

步骤2:上传华为Ascend模型

和上传英伟达GPU模型的操作类似，上传华为Ascend模型同样需要上传Tron model的相关信息，系统会根据下面填写的信息按照一定规则组合在顶部“模型文件名称”这一栏自动生成Tron model的模型文件名称。

* Repo：模型的类别（必选项，展示默认值）
* 模型名称：用户自行命名的模型名称（必填项，对输入内容没有限制）
* method（模型的网络格式）：模型的推理后处理方法（必选项，展示默认值）
* 版本：模型的版本信息（必填项，对输入内容没有限制）
* 模型文件：在这里上传的Tron model文件。（必填项，对上传文件格式不进行校验，请上传tron model文件）
* 选择模版：该Tron model是由哪个模型转换模版转换得来的（非必选项，模型转化模版在模型转换模块的模型转化页面，如未找到使用的转换模版，请上传转换模版后再上传tron model）
* 说明：对这个模版的描述，会在tron model列表中展示（非必填项，对输入内容没有限制）

image

image

上文中在英伟达GPU硬件平台的Tron模型管理中，提到过这些模型进的操作方式，这里就不再做过多的赘述； :::info 注意 目前系统尚未支持对华为Ascend模型的量化功能； :::

# 2.模型结构管理

在 Ti 深度学习平台中，基于Pytorch深度学习框架，输出的模型主要是模型权重的优化，此处PTH模型只保存了模型的操作符和具体的参数值，没有保存拓扑结构；为了让模型可以在英伟达GPU或者华为Ascend盒子硬件平台上运行推理测试任务，算法人员需要将PTH模型进行参数压缩和运算方式定义，并进行模型结构封装，所以在PTH模型进行转换之前需要创建对应的模型结构；

下面会详细介绍如何在 Mo 上如何创建模型结构： 在进行模型转化时，用户如果有新模型结构需要转换，且已有模型结构无法满足需求时，可以自助添加新模型结构进行转换。完成模型转换的模型结构会自动加入到“已认证”的模型结构。

步骤1:进入“模型转换”——在二级菜单中点击进入“模型结构管理”，在“未认证”tab页面点击“新建” image

步骤2:进入添加模型结构页面，输入模型结构相关信息；

* 模型结构名称：必填项，100个字符内；
* 模型结构类名：必填项，100个字符内；
* 类型配置文件：选择这一模型结构对应的类型配置文件，即为自己上传的不同版本网络模型（必选项）；
* 模型结构文件：上传模型结构文件；

点击“确认”后，完成新的模型结构的添加；

:::info 注意

py文件最后一个函数的注解里写的名字要和模型结构类名一致, 而且不能和已有的结构类名重复；

\*\* 如果出现重复, 需要修改py文件！\*\*

:::

image

image

模型结构认证：创建模型结构完成后，该模型结构属于未认证的模型结构，直到用户使用该模型结构成功完成一次模型转换后，该模型结构会被系统自动归类到“已认证”列表中

image

image

# 3. 模型转换模版

由于pth模型转换成tron模型有很多模型结构，以及压缩参数等配置需要用户配置，用户创建转换任务比较繁琐，因此系统通过模型转换模版来简化模型转换任务创建的操作步骤，用户可以按照自己的模型的参数新建一个转模型的参数模板，以后需要转模型，直接使用该模板保存的参数或稍加修改即可，具体操作如下：

步骤1:查看已存在的转换模版：进入模型转换——模型转换模版这一栏目，可以看到系统中自己创建的以及所有的转换模版；  
如果与自己需要转换的模型参数一致可以直接使用该模版进行模型的转化，否则需要重新创建新的转换模版。右上角搜索栏支持对转换模版的ID和模版的名称进行模糊搜索。

image

image

步骤2:上传新的转换模版：在模型转换——模型转换模版这里栏目，进入“我的”tab页面，点击“添加模版”

image

image

步骤3:输入新转换模版的参数

主模版：

* 模版名称：转换模版的名称（必填项，名字不能与其他模版重复）
* repo：用户转换模型的类别（必须选项）
* version：转换的起始版本号。
* method：这个是模型的推理后处理方法，目前是由数据库配死，分类模型就选择classification，检测模型需要根据自己转换模型的情况进行设置。
* error\_threshold: pytorch 和 tronmodel 推理的误差阈值（必填项，支持科学记数法：1e-5，小数法：0.00005）

image

image

子模版：（如果这个模型是由多个模型封装而成，请新增并填写子模版信息）

* 子模版名称：命名该子模版（必填项，100个字符内）；
* modelType：转模型的类型，一般都是pytorch模型（必选，数据库配死）；
* input\_name：例如：raw\_data（必填，200个字符内,不要填写括号）；
* input\_shape:输入的一批图片的size，  格式为（batch size，颜色通道，h，w）（必填1000个字符内，使用逗号将每个字符隔开即可，不要填写括号）；
* meanValues：归一化的均值（必选，不要填写括号）；
* scaleValues：归一化的方差（必选，不要填写括号）；
* maxbatch:最大的样本数量（必填，字符类型unit）；
* architecture：选择模型结构，就是之前提到的backbone（必选）；
* stateDicKey：模型参数存放dict的key名称，在自己save模型时设置的（必选）；
* num classes：分类数（必填）；

image

image

分类任务信息：

class\_name:分类名称。每一个维度表示一个分类任务，每一个维度里的分类数量就是该分类任务的分类维度。用户可以进行两种模式的编辑：

* 视图模式——用户通过交互的方式进行分类的编辑；
* code模式——用户直接通过编辑数组来实现编辑的功能；

args：额外的参数，该 参数是传入到推理引擎的参数，比如阈值、achor输出blob\_name等参数。分为vi（int的 vector）si（int）等几种类型（非必填项，使用逗号将每个字符隔开即可，不要填写括号）；

image

image

步骤4:完成转换模版的创建：点击“确认”完成转换模版的配置。

步骤5：查看转化模版：完成创建转换模版后，可以在模型转换模版这一页面的“我的”以及“全部”这两个tab页看到创建的转换模版，可以使用该模版创建转换任务。

# 4.模型转换任务

在 Ti 深度学习系统中基于Pytorch框架训练获得模型，训练平台输出的通常是权重的优化，几乎没有训练会使用模型的网络结构变化的策略，例如剪枝，等；等待训练执行到设置好的epoch会输出的模型的评估和存储下来的模型文件，存储下来的模型就是PTH模型；

我们这里说的PTH模型只保存了模型的各个操作符和具体的参数值，没有保存拓扑结构；

Tron模型是根据PTH模型通过压缩参数，和一些运算方式，并保存了拓扑结构可以得到Tron模型；

有两个硬件平台需要运行tron模型  
（1）运行在英伟达 GPU硬件上的Tron模型：由PTH模型直接转换得到的Tron模型可以直接创建模型推理任务，运行的硬件平台在英伟达GPU上；

（2）运行在华为Ascend盒子上的Tron模型;

算法用户需要将模型在上述两个硬件平台跑推理测试任务，就需要将模型转换成对应的模型格式；

这一章节我们会详细介绍如何使用 Mo 转换PTH模型和TRON模型；

# 4.1 PTH模型转TRON模型

PTH模型是没有办法直接用来跑推理测试任务的，因为PTH模型只包含了模型的操作符和参数值没有模型的结构；而运行推理测试的模型需要封装模型结构；

因此，我们将PTH转换成了Tron模型，里面封装了模型结构，操作符和执行参数，用于跑推理测试任务；转换好的Tron模型就可以在英伟达GPU上跑推理测试任务；

步骤1:从左侧导航栏中，点击进入“模型转换”——“模型转换任务”——“NVIDIA平台“ 在“我的”tab页面点击“创建任务”进入模型转换任务的配置页面； image

步骤2:选择需要转换的PTH模型、对模型转换的执行参数和使用的转换模版； [“转换模版”](/model-shop/instruction/transfer_template)

模型保存名称：模型保存名称会按照用来转换的PTH模型，以及创建转换任务的年月日的时间信息生成默认保存名称；

选择python模型：选择需要转换的PTH模型，模型为“PTH模型管理”中新建上传的PTH模型；

execution parameter：选择执行参数，可以选择拷贝模型参数或合并操作符，支持多选；

选择模版：选择创建好的转换模版，点击选择某个模版，则显示主模版信息和子模版信息，用户可根据实际模型转换的需要修改某些参数，其中； image

步骤3:点击确认后完成转换任务的创建，可以在模型转换任务列表中跟踪转换任务的进度；

注意：完成创建后任务会自动启动，无需用户手动操作； :::info 注意

完成创建后任务会自动启动，无需用户手动操作；

:::

# 4.2 不同硬件平台TRON模型转换

由于运行的平台不同，对模型的参数要求也不同，在英伟达GPU和华为Ascend盒子上，由于输入图片的标准不同，所以在转换成不同平台运行的模型时，输入图片的标准化的参数参数和配置有所不同，因此在模型转换的时候需要添加数据格式转换的操作；

步骤1:从左侧导航栏中，点击进入“模型转换”——“模型转换任务”——“Ascend平台“ 在“我的”tab页面点击“创建任务”进入模型转换任务的配置页面；

image

image

步骤2:选择需要转换的Tron模型，配置转换的执行参数

选择模型：选择上传或转换得到的NVIDIA平台的Tron模型；

色彩空间：由于模型推理需要输入并读取不同色彩空间存储的图片，所以我们也要对模型所输入的图片的色彩空间进行转换，目前支持下列的的图像通道转换： \* rgb转换成bgr

* bgr转换成rgb
* nv12转换成bgr
* nv12转换成rgb

mean\_values、scale\_values这两个参数是对模型读入数据进行归一化操作的两个参数配置，模型按照设定，对输入数据的预处理计算的系数；

* MeanValues：为整型数组，数组数据库配死；
* ScaleValues：float型数组，数组数据库配死；

步骤3:完成配置后，点击确认创建模型转换任务；

:::info 注意

完成创建后任务会自动启动，无需用户手动操作；

:::

# 5.模型量化

模型量化就是将浮点存储（运算）转换为整型存储（运算）的一种模型压缩技术。原来表示一个权重需要使用float32表示，量化后只需要使用int8来表示。

通过模型量化这一操作，我们可以对模型的网络进行加速，并且量化后模型的准确率和转化之前的相近或者误差在可以接受的范围内；

模型量化指的是对模型的权重、激活函数的输出和梯度进行量化，Mo 中的模型量化功能指的主要是对模型的权重进行量化，达到减少模型大小内存和占用空间的目的；

目前 Mo 支持对NVIDIA模型进行量化，下面就会详细介绍如何在系统中使用这一功能；

注意：系统仅支持对由PTH模型转换成的NVIDIA平台的Tron模型进行量化；

步骤1:点击进入“模型管理”——“Tron模型”——“NVIDIA平台”；

步骤2:找到需要量化的模型，点击操作栏中的“量化”按键，进入模型量化任务的配置页面；

步骤3:配置模型量化任务

* 模型ID：模型量化后还是原先的模型，因此模型ID不变，用户无需编辑；
* 模型名称：模型名称不变，用户无需编辑；
* batch大小：一次将测试集塞入到模型网络中样本的数量，必填项；
* calibration type：校准类型，为模型量化的策略，为下拉单项选择，int型，默认值为2；  
  取值范围为【0，1，2，3】，必选项；
* 是否使用测试集：关闭这一选项，需要上传图片作为模型量化的验证数据，打开这一选项，需要在下方选择数据集，必填项；
* 预处理：可以选择以下五种预处理方式对输入的验证数据：  
  【resize，centercrop，equal\_resize，cvtRGB，transpse】

image

image

# 6.数据集管理

模型推理测试任务需要使用数据集创建任务，因此不同版本的模型对这个数据集的推理结果通过精度、模型网络性能等指标的横向对比，直观的看出不同版本模型的性能和进度的提升等情况；  
模型推理和量化使用的数据集也是在这个模块维护，用户在创建任务前也要提前准备好推理任务和模型量化任务需要使用的数据集；  
本章会介绍如何导入DataOcean中的数据集进行创建和数据集管理模块的使用姿势；

## 6.1 新建数据集

在数据集管理的“我的”tab页，可以选择dataocean数据工作台的数据集进行创建，点击“新建”，进入添加数据集页面： image

步骤1：输入数据集名称，必填项，全局唯一；  
步骤2：选择来自dataocean-数据工作台的数据集，必选，基本固定的几种类型，与「模型转换模板」中的后处理method有映射对应关系；  
步骤3:说明：非必填项，限制字符数量 1000；  
步骤4:配置分类任务信息： - 选择关联模型转换模版和子模版； - 左侧：系统自动提取相关的分类任务的信息，用户不可直接编辑分类； - 右侧：用户根据模板的分类，配对数据集携带的标注信息：若数据集类型是【classify】：则罗列所有该数据集对应标注信息下classifications属性下的key（仅限tool值为radio的key）；若数据及类型是【detection】：则把所有key作为分类形成一个分类任务，此时一个数据集只有一个分类任务；  
注意：所有已经使用过的key，不可再选中；

:::info 注意

* 当模板的任何一个分类任务下的分类信息与对应的数据集的标注信息中的分类信息不匹配时，系统自动提示「数据不匹配，系统将以模版信息对数据进行处理」，用户点击确定后开始数据集创建；
* 在处理数据集时，任何一条数据的任意一个分类任务中的分类值不属于对应的模版分类任务分类集时，该条数据会被直接丢弃！

::: ## 6.2 数据集管理

在数据集管理的“全部”tab页，可以查看所有数据集的详情，进行预览，查看推理和下载的操作。

image

image

操作说明：

预览：查看数据集中每张图片的原图、渲染图以及json文件。

image

image

查看推理：查看使用了这个数据集的推理任务的输出结果的详情； 可以看到所用使用这个数据集运行推理任务的推理任务ID，数据集名称，数据集类型，推理模型，推理任务状态以及模型的精度，推理的总时长，预处理时长，推理时长，后处理时长以及推理任务的创建时间。

这样就可以横向对比不同版本模型的精度和性能，达到模型推理测试的目的；

image

image

下载：用户可以下载该数据集到本地；

查看批量结果：选中需要查看推理结果的任务，点击“查看批量结果”，即可查看多个推理测试任务的推理结果；  
批量结果可以查看多个模型对该数据集的推理详情、推理参数、结果评估和PR曲线图（检测任务）或者PR柱状图（分类任务）

image

image

## 6.3 数据集处理

在使用数据集的过程中，如果需要对数据集进行一些合并筛选的处理，包括但不限于合并某些数据类别，去掉某些数据类别等，可以使用[dataocean-python-sdk](/hc/hubble/Dataset/python_sdk)工具进行处理，更灵活更自由，满足各种数据集处理的需要。

# 7.模型推理测试任务

此功能模块主要包括两个内容，第一是对转换好的模型创建推理测试任务，第二个是对推理结果进行评估，查看推理测试任务的推理精度和推理任务的不同指标；

## 7.1 新建推理测试任务

这一部分会详细介绍如何创建一个推理测试任务；

步骤1:在模型推理——模型推理测试——我的，栏目中点击“新建”，模型推理测试任务会按照任务的更新时间进行倒序排列。

image

image

step2:配置推理测试任务的参数和信息，如下图所示

（1）选择模型：选择转换好的Tron model（必选）

（2）快速验证：开关默认关闭。当开关关闭时，用户可以选择对应的数据集进行推理，一次选择一个数据集；当开关开启时，用户可以从本地上传对应的图片来进行快速推理验证，一次可以选择多个图片。推理的结果存储在任务中，但是不会与数据集产生关联。

（3）选择数据集：选择通过DataOcean导入到 Mo 的数据集进行推理任务。现在是只有用1920乘1080的才可以保证新老版本的结果是一样，老系统好像把所有的图片，都默认1920乘1080P的，然后根据图上检测框的比例，去判定大目标小目标中目标，新系统是取图片的真实大小的。（必选）

（4）batch size：每次推理多大的batch，推理任务的最大样本总数。（必填）

（5）循环次数：循环推理次数，一般用于单图测速；（必填）

（6）评估IOU阈值：评估任务在数据集中检测相应物体准确度的一个标准；（必填）

（7）fp16和int8选择类型：前端根据后端的信息进行相应的显示：  
如果模型未经量化，则fp16和int8两个选项为disable的状态，不能选择；

如果模型经过量化，且最近一次是fp16，则fp16的按钮默认是关闭状态，可以操作；int8按钮是关闭状态，但是disable不可操作；

如果模型经过量化，且最近一次是int8，则fp16的按钮默认是关闭状态，可以操作；fp16按钮是关闭状态，但是disable不可操作；

image

image

（8）分类任务信息：会根据选择的模型中的分类信息自动匹配。

（9）统计维度：在实际业务中，用户需要将数据根据不同的维度分别进行统计，比如对人头人体的召回，需要区分白天/黑夜，或者晴天/阴天，借此可以了解模型在不同环境下的具体表现能力，从而可以有针对性的进行对应场景的提升。

为此，我们在业务上指导用户在 LabelX 中以分类属性的方式对数据进行标注，从而该类维度信息从广义上讲，也是数据标签的一部分，但是该标签不一定参与模型的推理。在MS创建推理任务时，我们要求用户将该类用于统计分析的维度标签也一并在【分类任务信息】中进行描述，然后再选择统计维度中进行选择，规范如下：  
1.用户可以点击“+”添加统计维度；  
2.每一个统计维度对应的分类任务默认关闭【开启为任务】，用户可以手动打开，此时推理任务将作为项目运行推理；  
3.用户在【分类任务信息】中删除对应的分类任务后，如果该任务被选为统计维度，则自动删除该统计维度；

（10）预处理：选择对数据集的预处理的方式和对应指标，支持添加多个，目前支持：resize、centercop、equal\_resize、cvt\_RGB以及transpose；

image

image

## 7.2 查看推理测试结果

步骤1:在推理测试任务中，可以点击查看运行完成任务的推理结果，点击操作栏中“查看结果”，进入推理结果查看页面；

image

image

可以看到推理详情、推理任务的具体执行参数、对不同目标的推理结果以及PR曲线图或PR柱状图；

image image