

Atlas 200I DK A2 开发者套件 23.0.RC3

转换模型

文档版本 01
发布日期 2023-11-14



版权所有 © 华为技术有限公司 2023。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编： 518129

网址： <https://www.huawei.com>

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

安全声明

漏洞声明

华为公司对产品漏洞管理的规定以“漏洞处理流程”为准，该政策可参考华为公司官方网站的网址：<https://www.huawei.com/cn/psirt/vul-response-process>。

如企业客户须获取漏洞信息，请访问：<https://securitybulletin.huawei.com/enterprise/cn/security-advisory>。

目录

1 内容介绍	1
2 在开发者套件上转换模型	2
3 在 Ubuntu 系统上转换模型	6
3.1 说明	6
3.2 安装方案	6
3.3 在虚拟机安装 Linux Ubuntu 22.04	7
3.3.1 安装 VMware Workstation 17 Player	7
3.3.2 使用 VMware 安装 Ubuntu 系统	8
3.3.3 安装 CANN	15
3.3.4 配置 SSH 文件传输	16
3.3.5 使用 ATC 命令转换模型	19
3.4 使用 WSL 安装 Linux Ubuntu 22.04	23
3.4.1 安装 Linux Ubuntu 22.04 子系统	24
3.4.2 安装 CANN	26
3.4.3 子系统的文件传输	26
3.4.4 使用 ATC 命令转换模型	27
3.5 在 PC 安装 Linux Ubuntu 22.04	30
3.5.1 安装 Ubuntu 22.04 系统	31
3.5.2 安装 CANN	38
3.5.3 双系统之间的文件传输	40
3.5.4 使用 ATC 命令转换模型	41
3.6 FAQ	45
3.6.1 运行 wsl --status 报错	46
3.6.2 运行 wsl --update 报错	48
3.6.3 无法解析服务器的名称或地址	48
3.6.4 在 PC 安装 Linux Ubuntu 22.04 后无法启动	49

1 内容介绍

对于开源框架的网络模型（如Caffe、TensorFlow等），不能直接在昇腾AI处理器上运行推理，需要先使用ATC（Ascend Tensor Compiler）工具将开源框架的网络模型转换为适配昇腾AI处理器的离线模型（*.om文件）。

当前支持在开发者套件和Ubuntu系统上转换模型，用户可以在开发者套件上进行模型转换，如果在开发者套件转换模型导致长时间不能运行（1小时以上），则请按照[3.2 安装方案](#)介绍的方法在Ubuntu系统上转换模型。

2 在开发者套件上转换模型

准备模型

步骤1 使用MobaXterm远程连接工具，以root用户登录开发者套件（密码：Mind@123）。

步骤2 获取网络模型，将下载后的压缩包上传至开发者套件任意目录，例如：“/home”。

- ONNX：单击[Link](#)或使用wget命令，解压压缩包，从“model”文件夹中获取*.onnx格式模型文件。

```
wget https://ascend-repo.obs.cn-east-2.myhuaweicloud.com/Atlas%20200I%20DK%20A2/DevKit/models/sdk_cal_samples/unetplusplus_sdk_python_sample.zip
```

- TensorFlow：单击[Link](#)或使用wget命令，获取*.pb格式模型文件。

```
wget https://ascend-repo-modelzoo.obs.cn-east-2.myhuaweicloud.com/c-version/ResNet50_for_TensorFlow/zh/1.7/m/ResNet50_for_TensorFlow_1.7_model.zip
```

- MindSpore：单击[Link](#)或使用wget命令，获取*.air格式模型文件。

```
wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com:443/cannInfo/model/resnet50_export.air
```

- Caffe：Caffe模型转换需要模型文件和权重文件。

- 模型文件（*.prototxt）：单击[Link](#)或使用wget命令下载该文件。

```
wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com/003_Atc_Models/AE/ATC%20Model/resnet50/resnet50.prototxt
```

- 权重文件（*.caffemodel）：单击[Link](#)或使用wget命令下载该文件。

```
wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com/003_Atc_Models/AE/ATC%20Model/resnet50/resnet50.caffemodel
```

步骤3 以下载ONNX模型为例，执行命令进入解压后模型文件存放目录。

```
cd /home/unetplusplus_sdk_python_sample/model
```

----结束

模型转换基础示例

以下介绍模型转换必须使用的参数。

步骤1 以ONNX模型为例，执行如下命令生成离线模型（如下命令中使用的目录以及文件均为样例），模型转换必须使用的参数介绍如[表2-1](#)所示。

```
atc --model=model.onnx --framework=5 --output=model --soc_version=Ascend310B4
```

📖 说明

转换模型失败，可参见[应用进程占用内存超出限制导致系统异常终止](#)解决。

表 2-1 参数说明

参数名	参数说明
--model	原始模型文件，填写模型文件时需要带上格式，如.onnx。
--weight	原始模型权重，该参数在转换Caffe模型场景下使用，其他框架不使用。
--framework	原始框架类型，各框架对应的数值如下： 0:Caffe; 1:MindSpore; 3:Tensorflow; 5:ONNX
--output	保存转换后的om离线推理模型文件路径。
--soc_version	昇腾AI处理器型号，填写“Ascend310B4”。

其中soc_version可使用npu-smi info命令查询当前芯片型号。

回显如下：

```
(base) root@davinci-mini:/home/ATC# npu-smi info
+-----+
| npu-smi 23.0.t30                               | Version: 23.0.t30                               |
+-----+-----+
| NPU  Name      | Health      | Power(W)  Temp(C)    Hugepages-Usage(page) |
| Chip Device    | Bus-Id      | AICore(%) Memory-Usage(MB) |
+-----+-----+
| 0   310B4      | OK          | 7.4       50         15 / 15              |
| 0   0          | NA         | 0         1309 / 3513         |
+-----+-----+
```

回显中加粗部分为芯片型号，填写参数时填写为“Ascend310B4”。

步骤2 若提示如下信息，则说明模型转换成功，若模型转换失败，则请参见[错误码参考](#)进行定位。

```
ATC run success
```

成功执行命令后，在--output参数指定的路径下，可查看离线模型（如：model.om）。

模型编译时，若遇到AI CPU算子不支持某种数据类型导致编译失败的场景，可通过启用Cast算子自动插入特性快速将输入转换为算子支持的数据类型，从而实现网络的快速打通，详细流程请参见[开启AI CPU Cast算子自动插入特性](#)。

----结束

其他常用转换参数说明

[模型转换基础示例](#)介绍了模型转换必须使用的参数，接下来以YoloV5模型和SVTR模型为例介绍其他常用参数的使用方法。更多模型转换参数说明请参见《[使用ATC工具转换模型](#)》中“参数说明”章节。

单击链接或使用wget命令下载[YoloV5模型](#)代码包，在model目录中获取onnx模型文件。

```
wget https://ascend-devkit-tool.obs.cn-south-1.myhuaweicloud.com/models/yolo_sdk_python_sample.zip
```

YoloV5模型转换命令如下：

```
atc --model=yolov5s.onnx --framework=5 --output=yolov5s_bs1 --input_format=NCHW --soc_version=Ascend310B4 --input_fp16_nodes="images"
```

单击链接或使用wget命令下**SVTR模型**代码包，在models目录中获取onnx模型文件。

```
wget https://ascend-devkit-tool.obs.cn-south-1.myhuaweicloud.com/models/ocr_acl_sample.zip
```

SVTR模型转换命令如下：

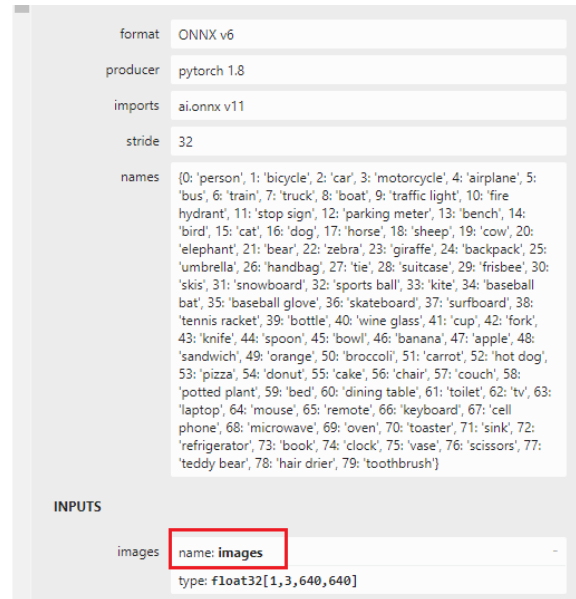
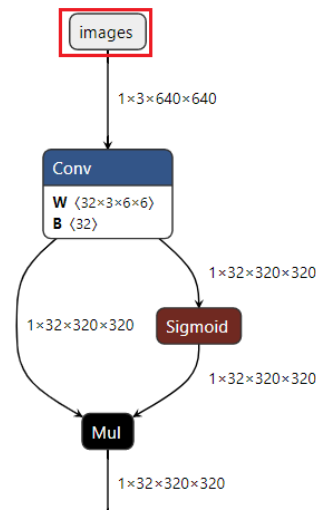
```
atc --model=svtr.onnx --framework=5 --input_shape='x:1,3,48,1440' --input_format=NCHW --soc_version=Ascend310B4 --output=svtr
```

表 2-2 参数说明

参数名	参数说明
--input_format	<p>输入Tensor的内存排列方式，NCHW指代batch、channels、height、width。</p> <ul style="list-style-type: none"> 当原始框架为Caffe时，支持NCHW、ND（表示支持任意维度格式，$N \leq 4$）两种格式，默认为NCHW。 当原始框架为ONNX时，支持NCHW、NCDHW、ND（表示支持任意维度格式，$N \leq 4$）三种格式，默认为NCHW。 当原始框架是TensorFlow时，支持NCHW、NHWC、ND、NCDHW、NDHWC五种输入格式，默认为NHWC。 <ul style="list-style-type: none"> 如果TensorFlow模型是通过ONNX模型转换工具输出的，则该参数必填，且值为NCHW。 如果原始模型中含有带data_format入参的算子，则该参数必填，推荐取值为ND，模型转换过程中会根据data_format属性的算子，推导出具体的format。若用户无法确定输入数据格式，则推荐指定为ND。 当原始框架为MindSpore时，只支持配置为NCHW。 <p>一般情况下不需要使用该参数，如果用户开发的应用代码前处理对内存排列有要求，可以使用该参数并填写所需的内存排列方式。</p>
--input_shape	<p>模型的输入节点名称和shape，shape的格式一般为[batch,channels,height,width]。</p> <p>一般情况下不需要使用该参数，如果要转换的模型为动态shape的ONNX模型时，需要使用该参数并填写shape。</p> <p>本文以将一个动态shape的SVTR模型转换为静态om模型为例。</p>
--input_fp16_nodes	<p>指定输入数据类型为FP16的输入节点名称。若不指定，则默认是float32数据类型。</p> <p>此参数可根据用户需要选择是否指定，若为默认float32，则精度相对较高；若为float16，则性能相对较高，在精度无明显下降的情况下有利于性能的提升。</p>

- 输入节点名称查看方法
使用**Netron模型可视化工具**打开（单击Open Model按钮）PC本地的模型文件，单击输入节点，查看输入节点名称。

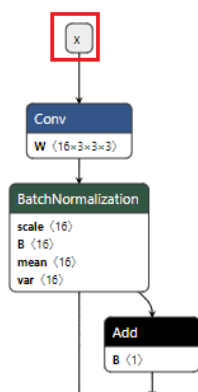
图 2-1 查看输入节点名称



- 输入节点shape查看方法和input_shape参数设置策略

使用Netron模型可视化工具打开模型，单击输入节点查看shape，可以看出输入节点x的shape为[p2o.DynamicDimension.0,3,48,p2o.DynamicDimension.1]，可以看出这个模型的输入是NCHW格式。第一个输入和最后一个输入分别为batch_size和width，且他们都由一个string填充，这种情况该维度为动态参数值（有时batch_size和width值为-1，这种情况和由string填充等价）。此时如果将该动态shape模型转换为batch size为1，宽度为1440的静态om模型时，input_shape参数可以设置为[1,3,48,1440]，1440为shape中的width，可以按需设置。

图 2-2 查看 shape



3 在 Ubuntu 系统上转换模型

- [3.1 说明](#)
- [3.2 安装方案](#)
- [3.3 在虚拟机安装Linux Ubuntu 22.04](#)
- [3.4 使用WSL安装Linux Ubuntu 22.04](#)
- [3.5 在PC安装Linux Ubuntu 22.04](#)
- [3.6 FAQ](#)

3.1 说明

对于开源框架的网络模型（如Caffe、TensorFlow等），不能直接在昇腾AI处理器上运行推理，需要先使用ATC（Ascend Tensor Compiler）工具将开源框架的网络模型转换为适配昇腾AI处理器的离线模型（*.om文件），开发者需要在个人PC或者服务器安装Ubuntu系统后，再安装CANN软件，使用ATC工具进行模型转换。

- 硬件资源推荐配置要求

表 3-1 硬件资源配置要求

资源	推荐起步规格
CPU	英特尔酷睿i5 4核
内存	16G

- 操作系统
推荐用户安装Ubuntu 22.04 LTS。

3.2 安装方案

在PC上安装Ubuntu 22.04 LTS有如下解决方案。

- 虚拟机方案：在Windows安装虚拟机，用虚拟机和镜像安装Ubuntu系统。
- WSL方案：使用适用于Linux的Windows子系统。
- 双系统方案：在PC上直接安装Ubuntu系统或Windows和Ubuntu双系统。

各个方案的优劣如表3-2所示，请读者按照自己的需求和条件进行选择。

表 3-2 各方案优劣势对比表

方案名称	优点	缺点
虚拟机	灵活	需要第三方虚拟化软件
WSL	Windows原生	受限于与微软的网络连接等问题
双系统	稳定	需要消耗一部分存储空间，且双系统需要来回切换

下面按照各方案详细介绍在PC使用ATC命令转换模型的流程。

3.3 在虚拟机安装 Linux Ubuntu 22.04

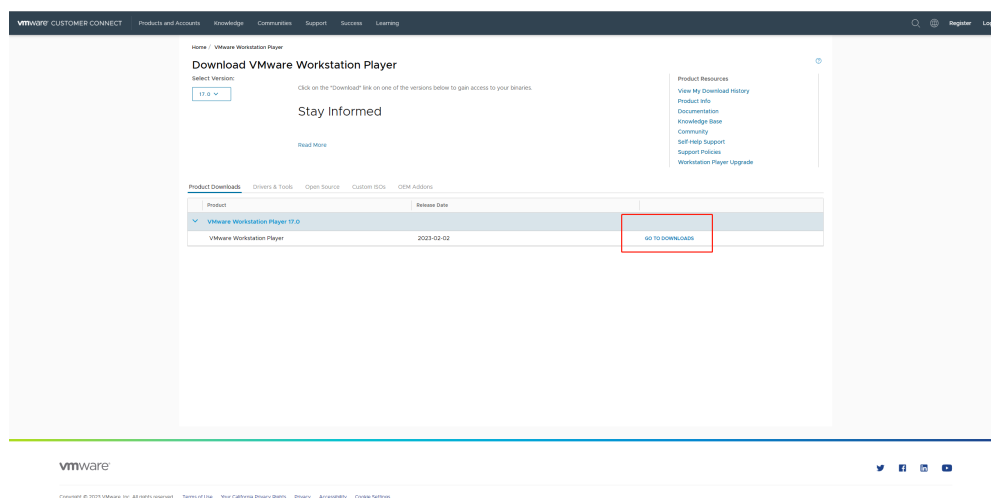
本章节介绍使用虚拟机安装Linux Ubuntu 22.04 LTS，并使用ATC命令转换模型的流程。

3.3.1 安装 VMware Workstation 17 Player

本节介绍如何下载和安装虚拟化软件VMware Workstation 17 Player。

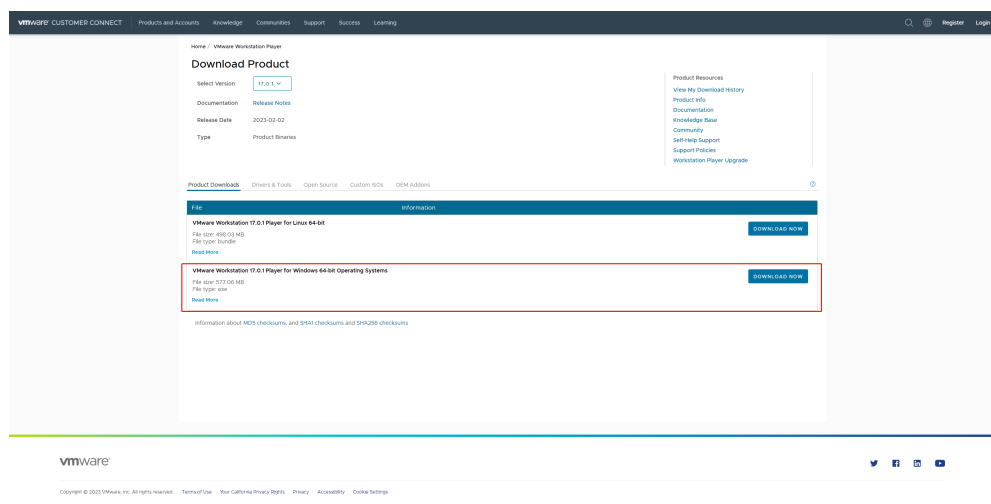
步骤1 单击[VMware Workstation 17 Player 下载页面](#)，按照图示进入下载链接。

图 3-1 VMware Workstation 17 Player 下载页面



步骤2 选择图示版本。

图 3-2 选择 Windows 64-bit 下载



步骤3 下载完成后按照工具的提示进行安装。

----结束

3.3.2 使用 VMware 安装 Ubuntu 系统

本节介绍如何下载Ubuntu 22.04 LTS镜像文件。

步骤1 单击[Ubuntu 22.04 LTS 下载链接](#)，按照图示版本将镜像下载到PC任意路径备用。

图 3-3 选择 Ubuntu 22.04.2 LTS 系统下载



步骤2 打开**3.3.1 安装VMware Workstation 17 Player**安装的VMware软件，选择“创建新虚拟机”。

图 3-4 创建新虚拟机



步骤3 选择“浏览”，选择步骤1中下载的镜像文件，单击“下一步”。

图 3-5 选择镜像文件



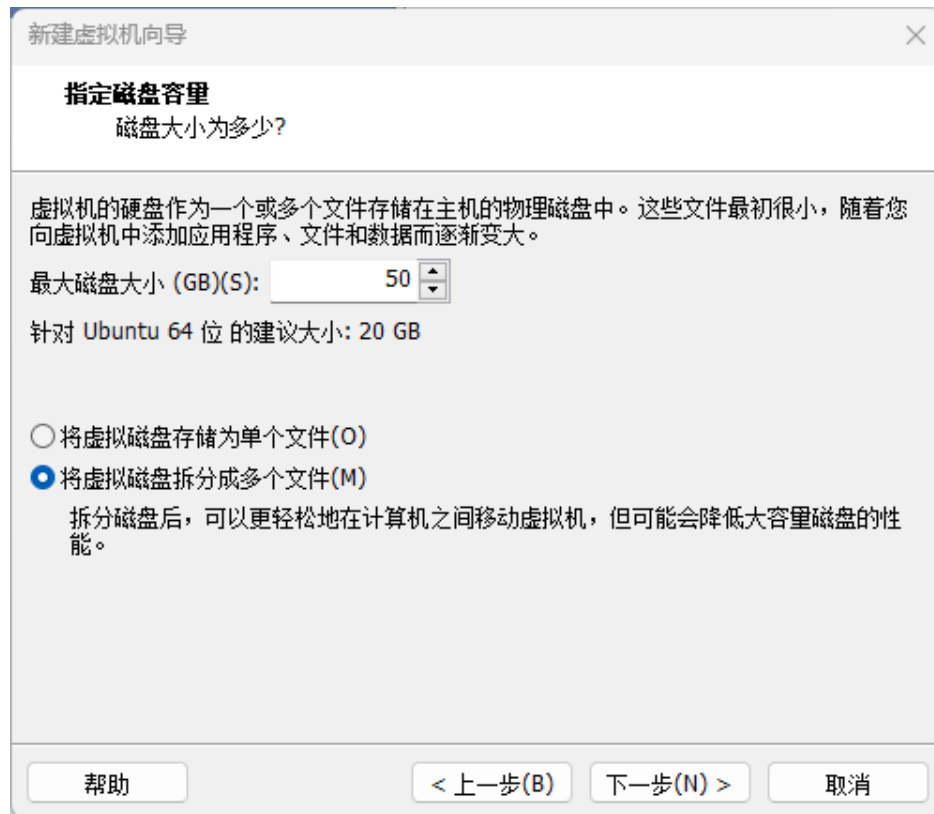
步骤4 自行设置用户名密码等，单击“下一步”。

图 3-6 个性化 Linux



步骤5 指定磁盘容量，此处建议至少分配50GB，选择“将虚拟磁盘拆分成多个文件”，单击“下一步”。

图 3-7 磁盘容量



步骤6 选择“自定义硬件”，为Ubuntu系统指定可以使用的内存和处理器核数，修改为16G内存与8核处理器。

图 3-8 自定义硬件

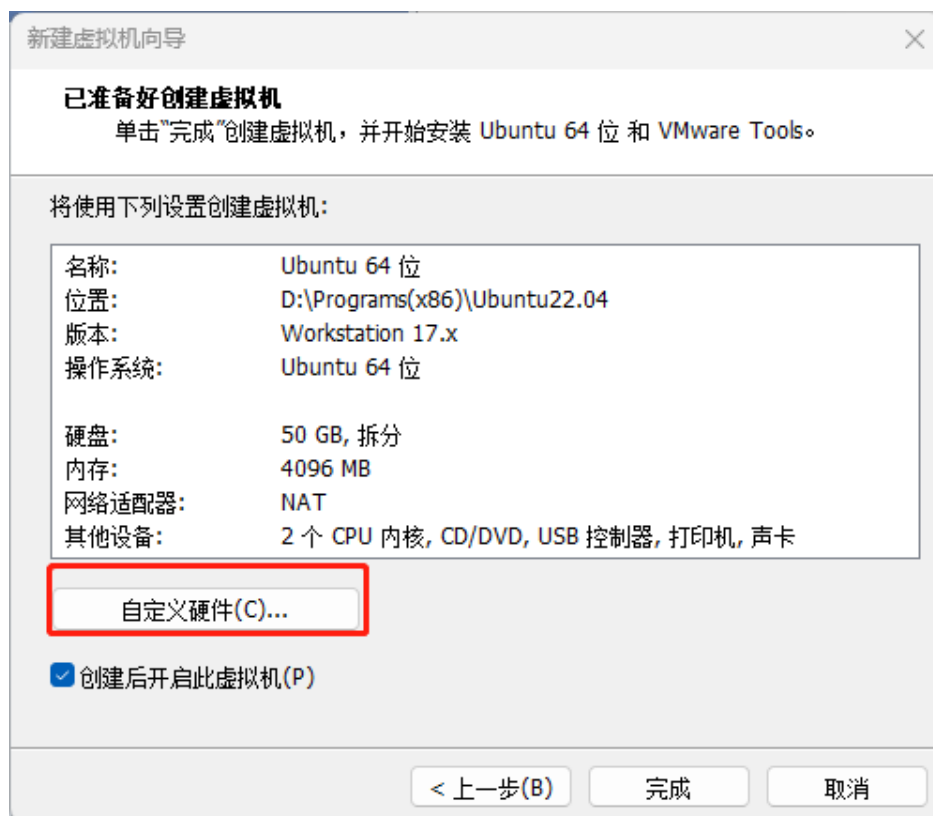
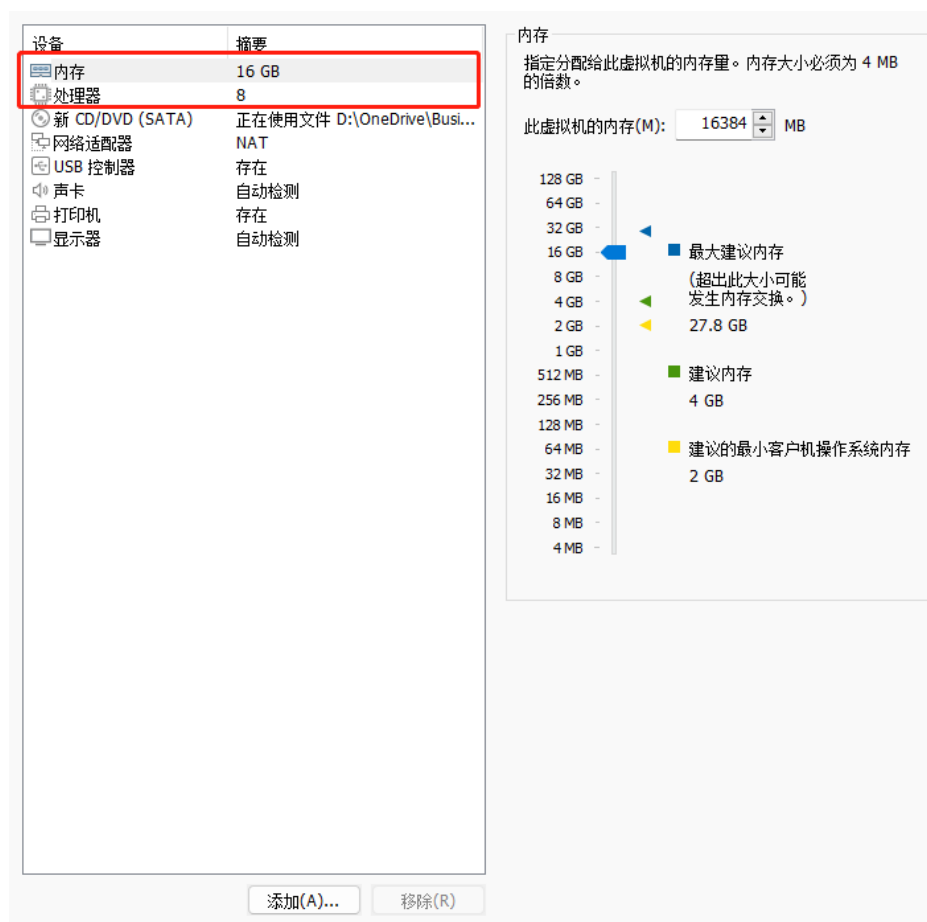
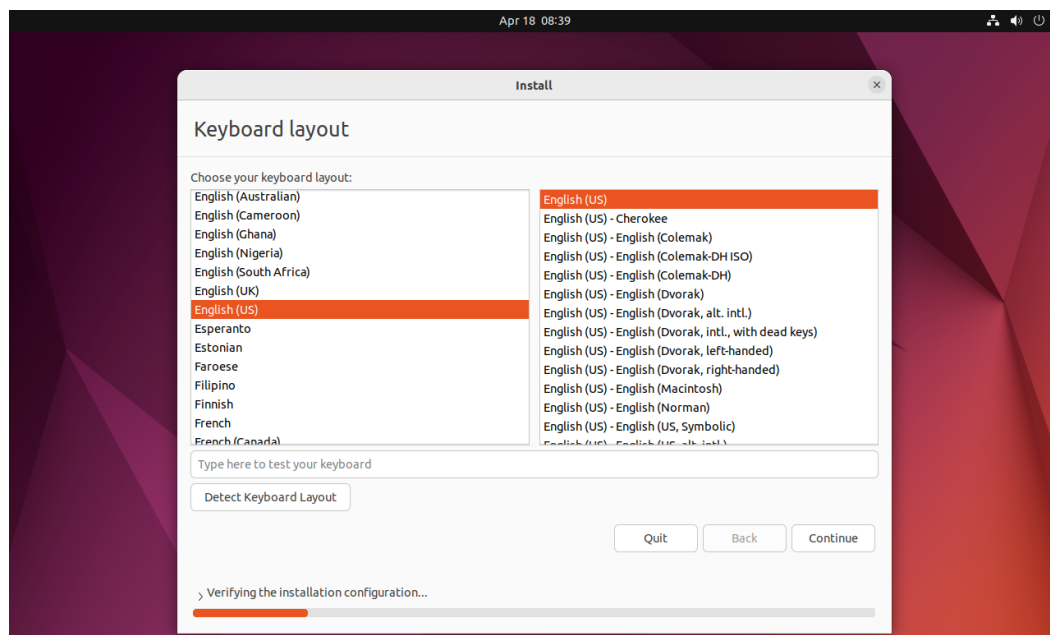


图 3-9 选择分配的内存和处理器核数



步骤7 配置完成后单击“完成”即可进入虚拟机开始系统安装。

图 3-10 安装系统界面



安装过程按照软件的提示进行安装即可，无需其他特殊配置，完成后会以**步骤4**创建的普通用户账号登录系统，安装过程可能比较缓慢，请用户耐心等待。

----结束

3.3.3 安装 CANN

本节介绍如何安装依赖和CANN。

步骤1 开启Windows和虚拟机之间的复制粘贴功能。

在程序菜单中找到Terminal或者使用键盘组合快捷键“Ctrl”+“Alt”+“T”打开终端，如**图3-11**和**图3-12**所示。

图 3-11 单击 LaunchPad

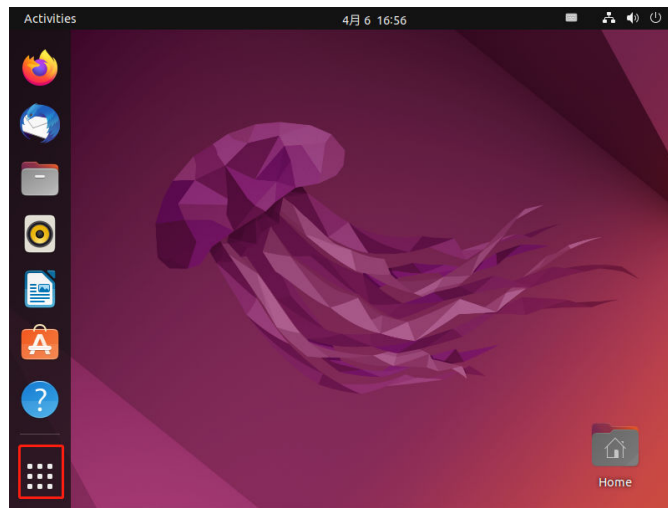
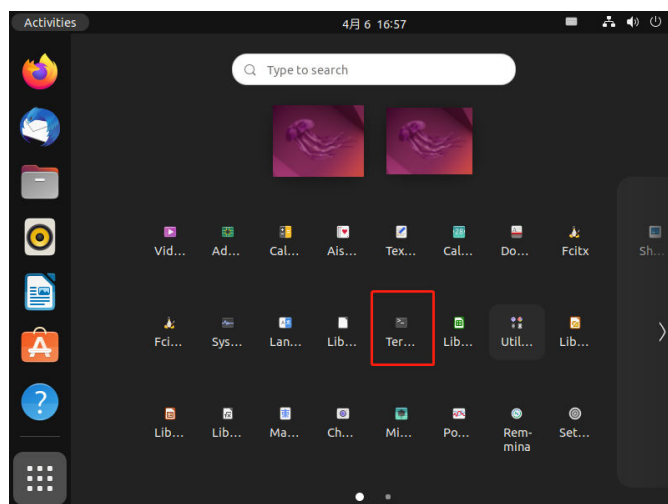


图 3-12 Terminal 工具



执行以下命令安装VMware Tools。

```
sudo apt install open-vm-tools-desktop -y
```

📖 说明

使用sudo命令时，需输入安装Ubuntu系统时创建的用户密码。

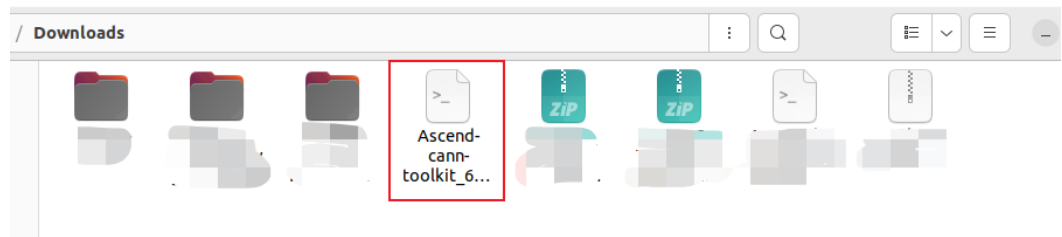
执行以下命令重启Ubuntu系统。

```
sudo reboot
```

步骤2 在Ubuntu浏览器中打开[下载链接](#)并下载CANN软件“Ascend-cann-toolkit_{version}_linux-x86_64.run”。

下载后文件会出现在“Downloads”目录中。

图 3-13 下载目录



📖 说明

一般情况下PC为x86架构，如果现场PC为Arm架构，请下载[Arm架构安装包](#)。

步骤3 参考[安装依赖](#)安装依赖。

步骤4 执行以下命令进行开发套件包的安装。

1. 进入“Downloads”目录并打开终端，增加对软件包的可执行权限。

```
chmod +x Ascend-cann-toolkit_7.0.RC1_linux-x86_64.run
```

2. 执行以下命令安装软件。

```
./Ascend-cann-toolkit_7.0.RC1_linux-x86_64.run --install
```

安装完成后，若显示如下信息，则说明软件安装成功：

```
[INFO] xxx install success
```

xxx表示安装的实际软件包名。

步骤5 安装完成后配置开发套件包的环境变量。

```
source CANN_INSTALL_PATH/ascend-toolkit/set_env.sh
export LD_LIBRARY_PATH=CANN_INSTALL_PATH/ascend-toolkit/latest/x86_64-linux/
devlib/:$LD_LIBRARY_PATH
```

CANN_INSTALL_PATH: 为CANN软件安装目录。

----结束

3.3.4 配置 SSH 文件传输

用户从PC windows系统上传待转换的模型文件到虚拟机Ubuntu系统，需要参见本节内容配置SSH文件传输。

步骤1 在虚拟机终端执行以下命令切换root用户。

```
sudo su root
```

回显如下：

```
[sudo] password for username:
```

输入密码后，命令行前缀切换为：

```
root@username-virtual-machine:
```

步骤2 安装所需依赖。

```
apt install firewalld openssh-client openssh-server
```

回显如下：

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
firewalld is already the newest version (1.1.1-1ubuntu1).
openssh-client is already the newest version (1:8.9p1-3ubuntu0.1).
openssh-server is already the newest version (1:8.9p1-3ubuntu0.1).
0 upgraded, 0 newly installed, 0 to remove and 100 not upgraded.
```

步骤3 开启虚拟机端口。

1. 添加端口。

```
firewall-cmd --permanent --add-port=22/tcp
```
2. 重新加载端口。

```
firewall-cmd --reload
```

步骤4 查询虚拟机IP。

```
ifconfig
```

回显如下：

```
ens33: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500
inet 192.168.xxx.xxx netmask 255.255.255.0 broadcast 192.168.xxx.xxx
inet6 fe80::bd01:3769:6c69:f842 prefixlen 64 scopeid 0x20
ether 00:0c:29:a0:12:d8 txqueuelen 1000 (以太网)
RX packets 48320 bytes 68416138 (68.4 MB)
RX errors 0 dropped 0 overruns 0 frame 0
TX packets 4021 bytes 377139 (377.1 KB)
TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0
```

步骤5 退出root用户权限。

```
exit
```

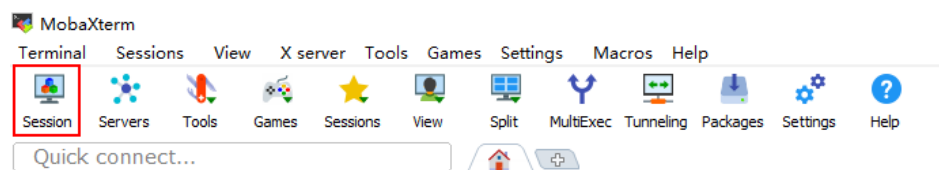
命令行前缀切换为：

```
username@username-virtual-machine:
```

步骤6 返回Windows系统，使用SSH远程登录软件（此处以MobaXterm工具为例）实现在PC Windows系统与虚拟机Ubuntu系统之间的文件传输，方法如下：

1. 单击打开MobaXterm进入主界面。
2. 单击左上方的“Session”进入界面。

图 3-14 单击 Session



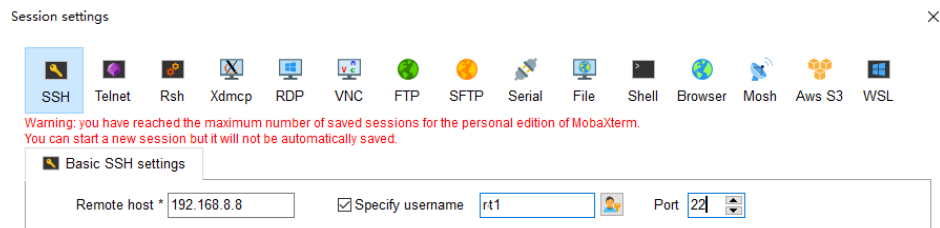
3. 单击左上方的“SSH”进入SSH连接配置界面。

图 3-15 单击 SSH



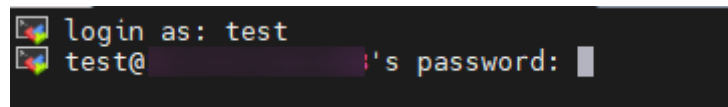
4. 配置Ubuntu系统用户和查询到的IP（以192.168.8.8为例，此处IP为步骤4中查询到的回显中加粗IP，请用户根据实际情况修改），单击“OK”按钮进入虚拟机Ubuntu系统。

图 3-16 填写登录信息



5. 输入Ubuntu系统的用户名与密码。

图 3-17 输入用户名与密码

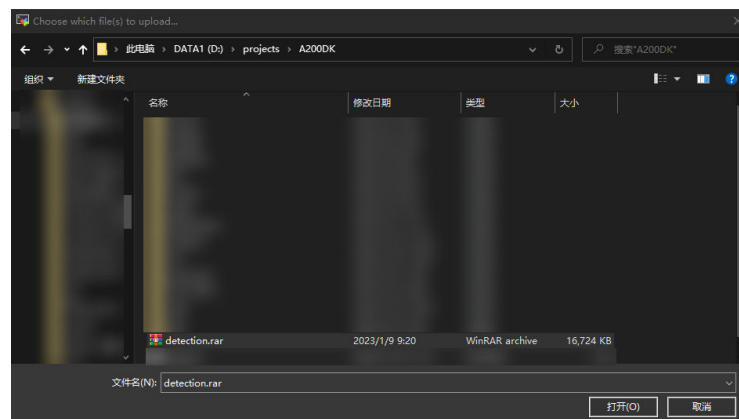


6. 上传与下载文件。
 - 上传文件。
单击上传按钮，如图3-18所示。选择待上传的文件，单击“打开”即可，如图3-19所示。

图 3-18 单击按钮



图 3-19 选择文件



- 下载文件。
单击待下载的文件，再单击下载按钮，如图3-20所示。选择文件在Windows中的下载位置，单击“OK”即可，如图3-21所示。

图 3-20 单击按钮

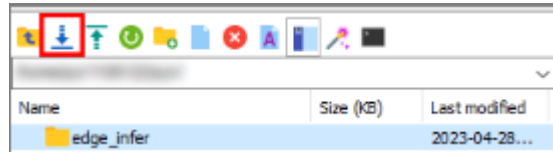
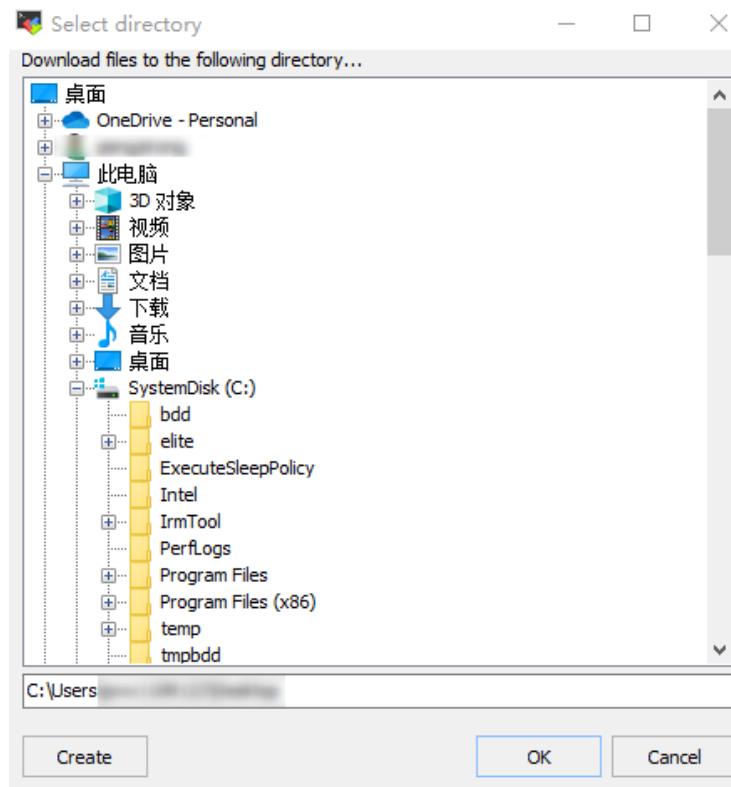


图 3-21 选择下载位置



----结束

3.3.5 使用 ATC 命令转换模型

本节介绍如何通过ATC工具将模型转换成支持在开发者套件上推理的离线om模型。

准备模型

步骤1 获取网络模型。

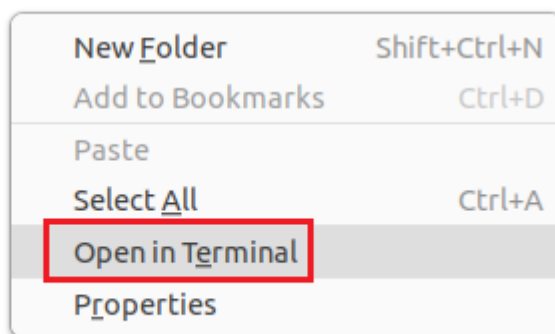
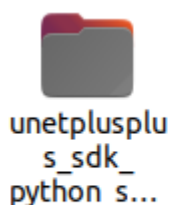
- ONNX: 单击[Link](#)或使用wget命令，解压压缩包，从“model”文件夹中获取*.onnx格式模型文件。
`wget https://ascend-repo.obs.cn-east-2.myhuaweicloud.com/Atlas%202001%20DK%20A2/DevKit/models/sdk_cal_samples/unetplusplus_sdk_python_sample.zip`
- TensorFlow: 单击[Link](#)或使用wget命令，获取*.pb格式模型文件。
`wget https://ascend-repo-modelzoo.obs.cn-east-2.myhuaweicloud.com/c-version/ResNet50_for_TensorFlow/zh/1.7/m/ResNet50_for_TensorFlow_1.7_model.zip`

- MindSpore: 单击[Link](#)或使用wget命令, 获取*.air格式模型文件。
wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com:443/cannInfo/model/resnet50_export.air
- Caffe: Caffe模型转换需要模型文件和权重文件。
 - 模型文件 (*.prototxt): 单击[Link](#)或使用wget命令下载该文件。
wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com/003_Atc_Models/AE/ATC%20Model/resnet50/resnet50.prototxt
 - 权重文件 (*.caffemodel): 单击[Link](#)或使用wget命令下载该文件。
wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com/003_Atc_Models/AE/ATC%20Model/resnet50/resnet50.caffemodel

步骤2 (可选) 如果是直接将模型文件下载到Ubuntu系统, 则可以直接进行模型转换操作; 如果是将模型文件下载到PC Windows系统, 需要参考[3.3.4 配置SSH文件传输](#)将模型上传到Ubuntu系统。

步骤3 在下载的代码文件存放目录右键单击空白处, 选择“Open in Terminal”打开命令行窗口。

图 3-22 打开终端



步骤4 以下载ONNX模型为例, 执行命令进入解压后模型文件存放目录。

```
cd unetplusplus_sdk_python_sample/model
```

----结束

模型转换基础示例

以下介绍模型转换必须使用的参数。

步骤1 以ONNX模型为例, 执行如下命令生成离线模型(如下命令中使用的目录以及文件均为样例), 模型转换必须使用的参数介绍如[表3-3](#)所示。

```
atc --model=model.onnx --framework=5 --output=model --soc_version=Ascend310B4
```


📖 说明

- 若用户在执行模型转换命令前关闭了配置环境变量的终端窗口，需参考[步骤3](#)与[步骤5](#)，在新终端窗口再次配置Python以及CANN的环境变量。
- 转换模型失败，可参见[应用进程占用内存超出限制导致系统异常终止](#)解决。

表 3-3 参数说明

参数名	参数说明
--model	原始模型文件，填写模型文件时需要带上格式，如.onnx。
--weight	原始模型权重，该参数在转换Caffe模型场景下使用，其他框架不使用。
--framework	原始框架类型，各框架对应的数值如下： 0:Caffe; 1:MindSpore; 3:Tensorflow; 5:ONNX
--output	保存转换后的om离线推理模型文件路径。
--soc_version	昇腾AI处理器型号，填写“Ascend310B4”。

步骤2 若提示如下信息，则说明模型转换成功，若模型转换失败，则请参见[错误码参考](#)进行定位。

```
ATC run success
```

成功执行命令后，在--output参数指定的路径下，可查看离线模型（如：model.om）。

模型编译时，若遇到AI CPU算子不支持某种数据类型导致编译失败的场景，可通过启用Cast算子自动插入特性快速将输入转换为算子支持的数据类型，从而实现网络的快速打通，详细流程请参见[开启AI CPU Cast算子自动插入特性](#)。

----结束

其他常用转换参数说明

[模型转换基础示例](#)介绍了模型转换必须使用的参数，接下来以YoloV5模型和SVTR模型为例介绍其他常用参数的使用方法。更多模型转换参数说明请参见《[使用ATC工具转换模型](#)》中“参数说明”章节。

单击链接或使用wget命令下载[YoloV5模型](#)代码包，在model目录中获取onnx模型文件。

```
wget https://ascend-devkit-tool.obs.cn-south-1.myhuaweicloud.com/models/yolo_sdk_python_sample.zip
```

YoloV5模型转换命令如下：

```
atc --model=yolov5s.onnx --framework=5 --output=yolov5s_bs1 --input_format=NCHW --soc_version=Ascend310B4 --input_fp16_nodes="images"
```

单击链接或使用wget命令下[SVTR模型](#)代码包，在models目录中获取onnx模型文件。

```
wget https://ascend-devkit-tool.obs.cn-south-1.myhuaweicloud.com/models/ocr_acl_sample.zip
```

SVTR模型转换命令如下：

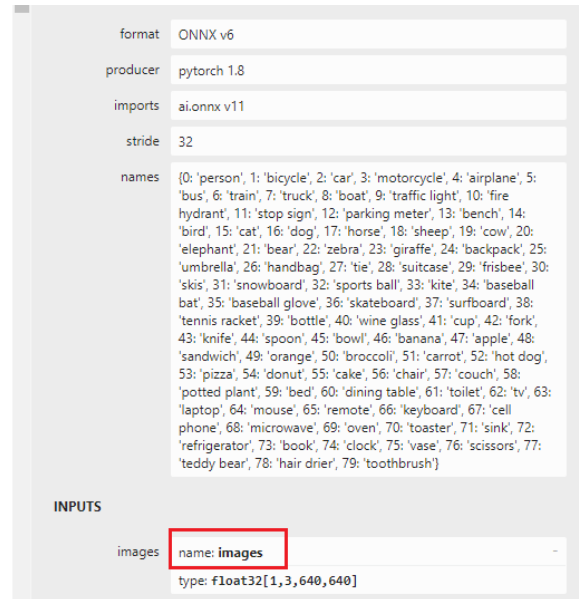
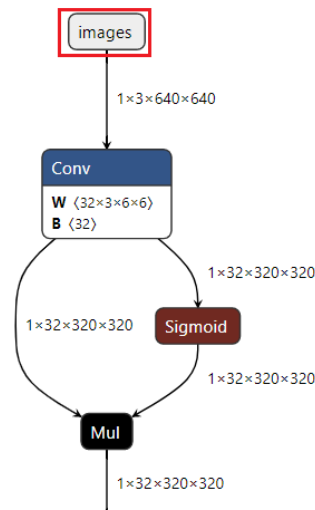
```
atc --model=svtr.onnx --framework=5 --input_shape='x:1,3,48,1440' --input_format=NCHW --
soc_version=Ascend310B4 --output=svtr
```

表 3-4 参数说明

参数名	参数说明
--input_format	<p>输入Tensor的内存排列方式，NCHW指代batch、channels、height、width。</p> <ul style="list-style-type: none"> 当原始框架为Caffe时，支持NCHW、ND（表示支持任意维度格式，N<=4）两种格式，默认为NCHW。 当原始框架为ONNX时，支持NCHW、NCDHW、ND（表示支持任意维度格式，N<=4）三种格式，默认为NCHW。 当原始框架是TensorFlow时，支持NCHW、NHWC、ND、NCDHW、NDHWC五种输入格式，默认为NHWC。 <ul style="list-style-type: none"> 如果TensorFlow模型是通过ONNX模型转换工具输出的，则该参数必填，且值为NCHW。 如果原始模型中含有带data_format入参的算子，则该参数必填，推荐取值为ND，模型转换过程中会根据data_format属性的算子，推导出具体的format。若用户无法确定输入数据格式，则推荐指定为ND。 当原始框架为MindSpore时，只支持配置为NCHW。 <p>一般情况下不需要使用该参数，如果用户开发的应用代码前处理对内存排列有要求，可以使用该参数并填写所需的内存排列方式。</p>
--input_shape	<p>模型的输入节点名称和shape，shape的格式一般为[batch,channels,height,width]。</p> <p>一般情况下不需要使用该参数，如果要转换的模型为动态shape的ONNX模型时，需要使用该参数并填写shape。</p> <p>本文以将一个动态shape的SVTR模型转换为静态om模型为例。</p>
--input_fp16_nodes	<p>指定输入数据类型为FP16的输入节点名称。若不指定，则默认是float32数据类型。</p> <p>此参数可根据用户需要选择是否指定，若为默认float32，则精度相对较高；若为float16，则性能相对较高，在精度无明显下降的情况下有利于性能的提升。</p>

- 输入节点名称查看方法
使用[Netron模型可视化工具](#)打开（单击Open Model按钮）PC本地的模型文件，单击输入节点，查看输入节点名称。

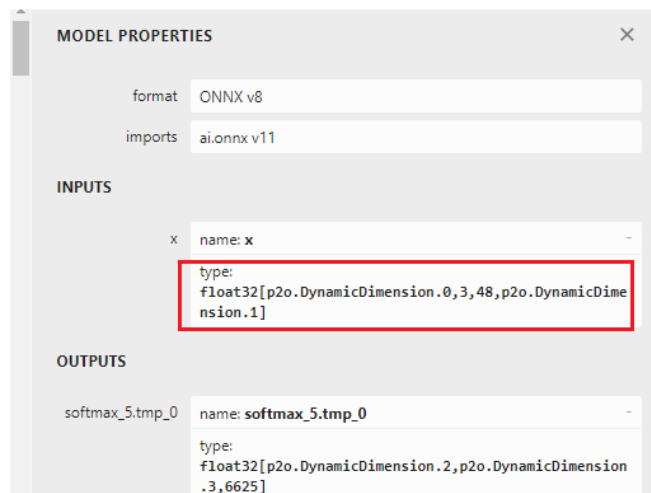
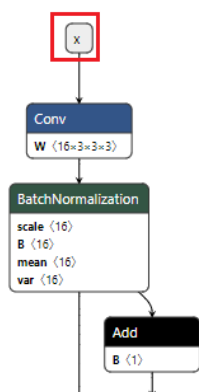
图 3-23 查看输入节点名称



- 输入节点shape查看方法和input_shape参数设置策略

使用Netron模型可视化工具打开模型，单击输入节点查看shape，可以看出输入节点x的shape为[p2o.DynamicDimension.0,3,48,p2o.DynamicDimension.1]，可以看出这个模型的输入是NCHW格式。第一个输入和最后一个输入分别为batch_size和width，且他们都由一个string填充，这种情况该维度为动态参数值（有时batch_size和width值为-1，这种情况和由string填充等价）。此时如果将该动态shape模型转换为batch size为1，宽度为1440的静态om模型时，input_shape参数可以设置为[1,3,48,1440]，1440为shape中的width，可以按需设置。

图 3-24 查看 shape



3.4 使用 WSL 安装 Linux Ubuntu 22.04

本章节介绍使用WSL安装Linux Ubuntu 22.04 LTS，并使用ATC命令转换模型的流程。

3.4.1 安装 Linux Ubuntu 22.04 子系统

以下步骤将详细介绍WSL在Windows上安装Linux Ubuntu22.04的操作过程，用户可参考[官方文档](#)进行子系统的安装与升级。

📖 说明

必须运行Windows 10版本2004及更高版本（内部版本19041及更高版本）或Windows 11才能使用以下命令。

步骤1 键盘输入“Win+R”，打开命令行窗口，输入cmd即可查看系统版本信息（在命令行输入VER也能达到相同效果），回显如下所示：

```
C:\Users\xxx>cmd
Microsoft Windows [版本 10.0.19045.2486]
(c) Microsoft Corporation。保留所有权利。
```

步骤2 搜索并找到Windows PowerShell/命令提示符，右键单击图标并选择“以管理员身份运行”后，先更新wsl内核，再通过运行以下命令查看在线商店可以下载并且可用的Linux发行版本列表：

```
wsl --update
wsl --list --online
```

返回结果需要时间，从几秒到几十秒不等，需要耐心等待。

图 3-25 发行版本列表

```
以下是可安装的有效分发的列表。
使用 'wsl.exe --install <Distro>' 安装。

NAME                                FRIENDLY NAME
Ubuntu                               Ubuntu
Debian                               Debian GNU/Linux
kali-linux                           Kali Linux Rolling
Ubuntu-18.04                         Ubuntu 18.04 LTS
Ubuntu-20.04                         Ubuntu 20.04 LTS
Ubuntu-22.04                         Ubuntu 22.04 LTS
OracleLinux_8_5                      Oracle Linux 8.5
OracleLinux_7_9                      Oracle Linux 7.9
SUSE-Linux-Enterprise-Server-15-SP4  SUSE Linux Enterprise Server 15 SP4
openSUSE-Leap-15.4                   openSUSE Leap 15.4
openSUSE-Tumbleweed                  openSUSE Tumbleweed
```

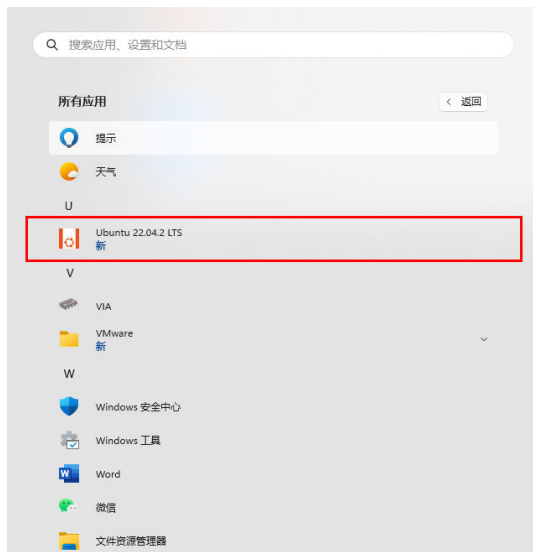
步骤3 确认需下载的Ubuntu 22.04版本，输入以下命令，安装完成后重启计算机。

```
wsl --install --distribution Ubuntu-22.04
```

步骤4 启动Linux Ubuntu22.04子系统。

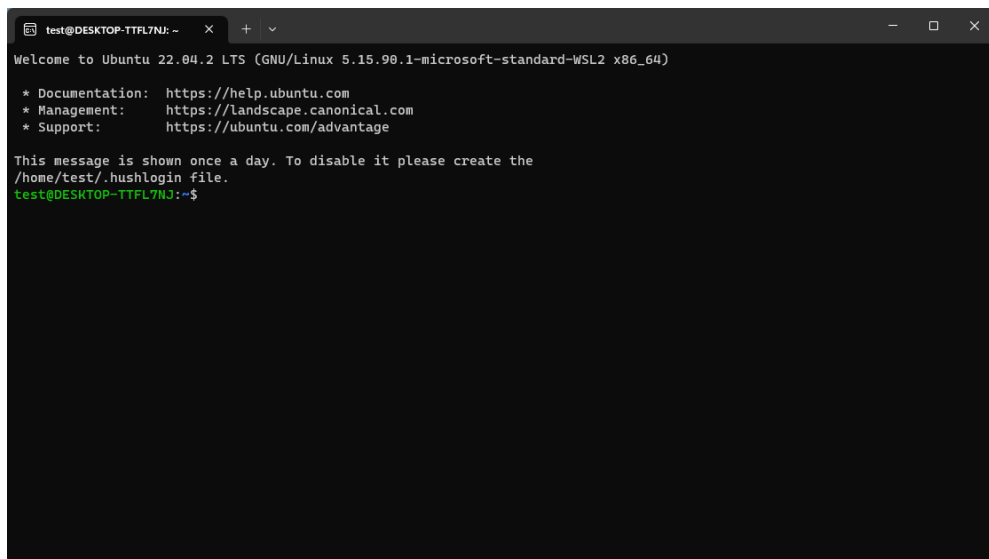
单击“Windows”菜单栏，所用应用下面找到并单击“Ubuntu 22.04.2 LTS”图标，启动Ubuntu 22.04.2 LTS，如下图所示。

图 3-26 所有应用下找到 Ubuntu 22.04.2 LTS



启动后，可以通过命令行界面操作Ubuntu 22.04 LTS系统，如下图所示。

图 3-27 Ubuntu 22.04 LTS 系统命令行操作界面



说明

首次启动Ubuntu 22.04 LTS系统时需要设置用户名与密码。

步骤5 在Ubuntu 22.04 LTS系统命令行界面配置apt-get源。

1. 备份原配置文件。
`sudo cp /etc/apt/sources.list /etc/apt/sources.list.bak`
2. 修改“sources.list”文件，将“archive.ubuntu.com”和“security.ubuntu.com”替换成“repo.huaweicloud.com”，请依次执行每行命令。
`sudo sed -i "s@http://.*archive.ubuntu.com@http://repo.huaweicloud.com@g" /etc/apt/sources.list`
`sudo sed -i "s@http://.*security.ubuntu.com@http://repo.huaweicloud.com@g" /etc/apt/sources.list`

3. 执行更新索引和更新软件命令并根据提示输入Yes。

```
sudo apt-get update  
sudo apt upgrade
```

📖 说明

使用sudo命令时，需输入安装Ubuntu系统时创建的用户密码。

----结束

3.4.2 安装 CANN

本节介绍如何安装CANN的依赖和CANN软件包。

- 步骤1 在浏览器中打开[下载链接](#)根据系统架构下载CANN软件“Ascend-cann-toolkit_{version}_linux-x86_64.run”。。

下载后的文件在Windows默认下载路径中对应Ubuntu子系统的路径为“/mnt/windows_default_download_path/Downloads”目录中，其中windows_default_download_path默认为“/c/User/username/”，username请根据实际情况替换。

📖 说明

一般情况下PC为x86架构，如果现场PC为Arm架构，请下载[Arm架构安装包](#)。

- 步骤2 参考[安装依赖](#)安装依赖。

📖 说明

使用WSL子系统时安装依赖时需使用sudo命令，否则会提示Permission denied。

- 步骤3 进入文件下载目录，执行以下命令进行开发套件包的安装。

1. 增加对软件包的可执行权限。

```
chmod +x Ascend-cann-toolkit_7.0.RC1_linux-x86_64.run
```

2. 执行以下命令安装软件。

```
./Ascend-cann-toolkit_7.0.RC1_linux-x86_64.run --install
```

安装完成后，若显示如下信息，则说明软件安装成功：

```
[INFO] xxx install success
```

xxx表示安装的实际软件包名。

- 步骤4 安装完成后配置开发套件包的环境变量。

```
source CANN_INSTALL_PATH/ascend-toolkit/set_env.sh  
export LD_LIBRARY_PATH=CANN_INSTALL_PATH/ascend-toolkit/latest/x86_64-linux/  
devlib:$LD_LIBRARY_PATH
```

CANN_INSTALL_PATH: 为CANN软件安装目录。

----结束

3.4.3 子系统的文件传输

Windows系统与Ubuntu子系统之间存在目录映射，Ubuntu子系统可直接访问Windows系统中的目录，其映射关系请参见[表3-5](#)。

若需要在系统之间进行传输，可直接在Windows系统中移动文件，再使用Ubuntu子系统访问对应目录。

表 3-5 系统目录映射

Windows系统盘目录	Ubuntu子系统映射目录
D:\	/mnt/d
C:\	/mnt/c

3.4.4 使用 ATC 命令转换模型

本节介绍如何通过ATC工具将模型转换成支持在开发者套件上推理的离线om模型。

准备模型

步骤1 获取网络模型。

- ONNX：单击[Link](#)或使用wget命令，解压压缩包，从“model”文件夹中获取*.onnx格式模型文件。
wget https://ascend-repo.obs.cn-east-2.myhuaweicloud.com/Atlas%20200i%20DK%20A2/DevKit/models/sdk_cal_samples/unetplusplus_sdk_python_sample.zip
- TensorFlow：单击[Link](#)或使用wget命令，获取*.pb格式模型文件。
wget https://ascend-repo-modelzoo.obs.cn-east-2.myhuaweicloud.com/c-version/ResNet50_for_TensorFlow/zh/1.7/m/ResNet50_for_TensorFlow_1.7_model.zip
- MindSpore：单击[Link](#)或使用wget命令，获取*.air格式模型文件。
wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com:443/cannInfo/model/resnet50_export.air
- Caffe：Caffe模型转换需要模型文件和权重文件。
 - 模型文件 (*.prototxt)：单击[Link](#)或使用wget命令下载该文件。
wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com/003_Atc_Models/AE/ATC%20Model/resnet50/resnet50.prototxt
 - 权重文件 (*.caffemodel)：单击[Link](#)或使用wget命令下载该文件。
wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com/003_Atc_Models/AE/ATC%20Model/resnet50/resnet50.caffemodel

步骤2 以下载ONNX模型为例，在Ubuntu系统命令行中执行命令进入模型存放目录。

Windows系统中“D:\Download\”路径在Ubuntu子系统上对应的“/mnt/d/Download/”文件夹。

```
cd /mnt/windows_default_download_path/Downloads
```

*windows_default_download_path*默认为“/c/User/username/”，*username*请根据实际情况替换。

步骤3 执行命令解压压缩包。

```
unzip unetplusplus_sdk_python_sample.zip
```

步骤4 执行命令进入解压后目录。

```
cd unetplusplus_sdk_python_sample/model
```

----结束

模型转换基础示例

以下介绍模型转换必须使用的参数。

- ##### 步骤1
- 以ONNX模型为例，执行如下命令生成离线模型（如下命令中使用的目录以及文件均为样例），模型转换必须使用的参数介绍如[表3-6](#)所示。

```
atc --model=model.onnx --framework=5 --output=model --soc_version=Ascend310B4
```

📖 说明

转换模型失败，可参见[应用进程占用内存超出限制导致系统异常终止](#)解决。

表 3-6 参数说明

参数名	参数说明
--model	原始模型文件，填写模型文件时需要带上格式，如.onnx。
--weight	原始模型权重，该参数在转换Caffe模型场景下使用，其他框架不使用。
--framework	原始框架类型，各框架对应的数值如下： 0:Caffe; 1:MindSpore; 3:Tensorflow; 5:ONNX
--output	保存转换后的om离线推理模型文件路径。
--soc_version	昇腾AI处理器型号，填写“Ascend310B4”。

步骤2 若提示如下信息，则说明模型转换成功，若模型转换失败，则请参见[错误码参考](#)进行定位。

```
ATC run success
```

成功执行命令后，在--output参数指定的路径下，可查看离线模型（如：model.om）。

模型编译时，若遇到AI CPU算子不支持某种数据类型导致编译失败的场景，可通过启用Cast算子自动插入特性快速将输入转换为算子支持的数据类型，从而实现网络的快速打通，详细流程请参见[开启AI CPU Cast算子自动插入特性](#)。

---结束

其他常用转换参数说明

[模型转换基础示例](#)介绍了模型转换必须使用的参数，接下来以YoloV5模型和SVTR模型为例介绍其他常用参数的使用方法。更多模型转换参数说明请参见《[使用ATC工具转换模型](#)》中“参数说明”章节。。

单击链接或使用wget命令下载[YoloV5模型](#)代码包，在model目录中获取onnx模型文件。

```
wget https://ascend-devkit-tool.obs.cn-south-1.myhuaweicloud.com/models/yolo_sdk_python_sample.zip
```

YoloV5模型转换命令如下：

```
atc --model=yolov5s.onnx --framework=5 --output=yolov5s_bs1 --input_format=NCHW --soc_version=Ascend310B4 --input_fp16_nodes="images"
```

单击链接或使用wget命令下[SVTR模型](#)代码包，在models目录中获取onnx模型文件。

```
wget https://ascend-devkit-tool.obs.cn-south-1.myhuaweicloud.com/models/ocr_acl_sample.zip
```

SVTR模型转换命令如下：

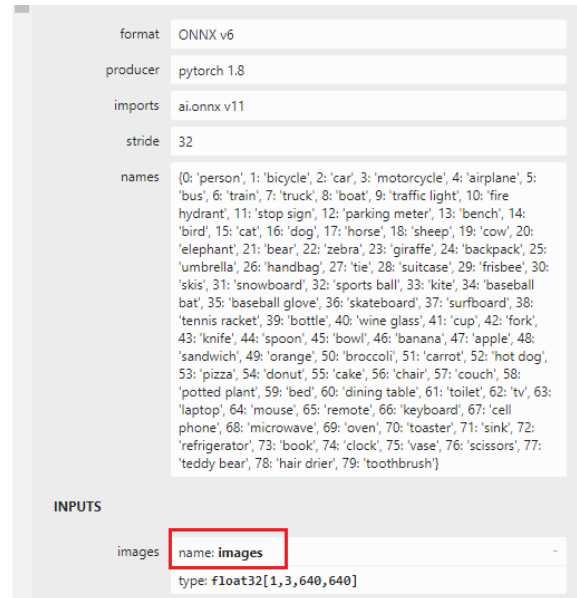
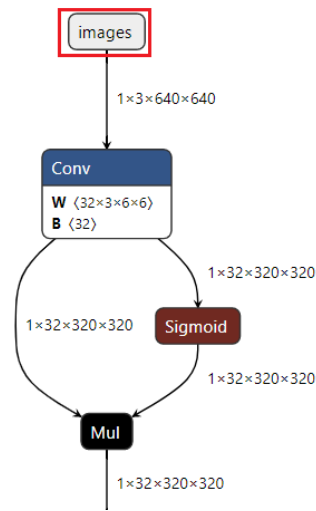
```
atc --model=svtr.onnx --framework=5 --input_shape='x:1,3,48,1440' --input_format=NCHW --soc_version=Ascend310B4 --output=svtr
```


表 3-7 参数说明

参数名	参数说明
--input_format	<p>输入Tensor的内存排列方式，NCHW指代batch、channels、height、width。</p> <ul style="list-style-type: none"> 当原始框架为Caffe时，支持NCHW、ND（表示支持任意维度格式，$N \leq 4$）两种格式，默认为NCHW。 当原始框架为ONNX时，支持NCHW、NCDHW、ND（表示支持任意维度格式，$N \leq 4$）三种格式，默认为NCHW。 当原始框架是TensorFlow时，支持NCHW、NHWC、ND、NCDHW、NDHWC五种输入格式，默认为NHWC。 <ul style="list-style-type: none"> 如果TensorFlow模型是通过ONNX模型转换工具输出的，则该参数必填，且值为NCHW。 如果原始模型中含有带data_format入参的算子，则该参数必填，推荐取值为ND，模型转换过程中会根据data_format属性的算子，推导出具体的format。若用户无法确定输入数据格式，则推荐指定为ND。 当原始框架为MindSpore时，只支持配置为NCHW。 <p>一般情况下不需要使用该参数，如果用户开发的应用代码前处理对内存排列有要求，可以使用该参数并填写所需的内存排列方式。</p>
--input_shape	<p>模型的输入节点名称和shape，shape的格式一般为[batch,channels,height,width]。</p> <p>一般情况下不需要使用该参数，如果要转换的模型为动态shape的ONNX模型时，需要使用该参数并填写shape。</p> <p>本文以将一个动态shape的SVTR模型转换为静态om模型为例。</p>
--input_fp16_nodes	<p>指定输入数据类型为FP16的输入节点名称。若不指定，则默认是float32数据类型。</p> <p>此参数可根据用户需要选择是否指定，若为默认float32，则精度相对较高；若为float16，则性能相对较高，在精度无明显下降的情况下有利于性能的提升。</p>

- 输入节点名称查看方法
使用[Netron模型可视化工具](#)打开（单击Open Model按钮）PC本地的模型文件，单击输入节点，查看输入节点名称。

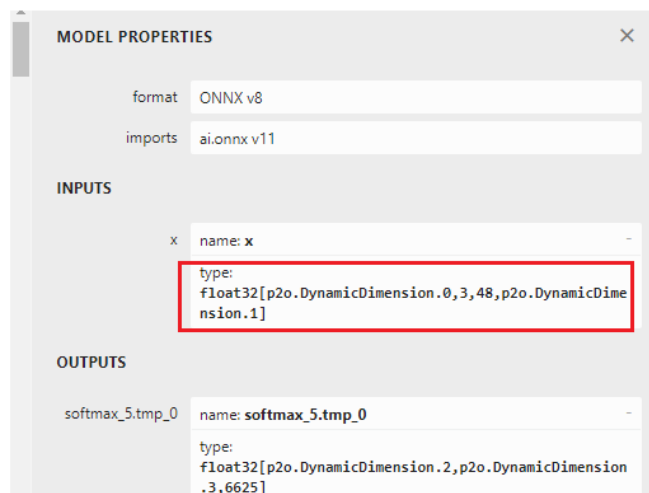
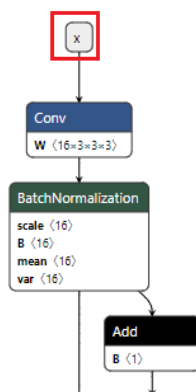
图 3-28 查看输入节点名称



- 输入节点shape查看方法和input_shape参数设置策略

使用Netron模型可视化工具打开模型，单击输入节点查看shape，可以看出输入节点x的shape为[p2o.DynamicDimension.0,3,48,p2o.DynamicDimension.1]，可以看出这个模型的输入是NCHW格式。第一个输入和最后一个输入分别为batch_size和width，且他们都由一个string填充，这种情况该维度为动态参数值（有时batch_size和width值为-1，这种情况和由string填充等价）。此时如果将该动态shape模型转换为batch size为1，宽度为1440的静态om模型时，input_shape参数可以设置为[1,3,48,1440]，1440为shape中的width，可以按需设置。

图 3-29 查看 shape



3.5 在 PC 安装 Linux Ubuntu 22.04

本章节介绍使用PC直接安装Linux Ubuntu 22.04 LTS，并使用ATC命令转换模型的流程。

3.5.1 安装 Ubuntu 22.04 系统

步骤1 单击开始菜单，搜索磁盘管理工具，选择“创建并格式化硬盘分区”。

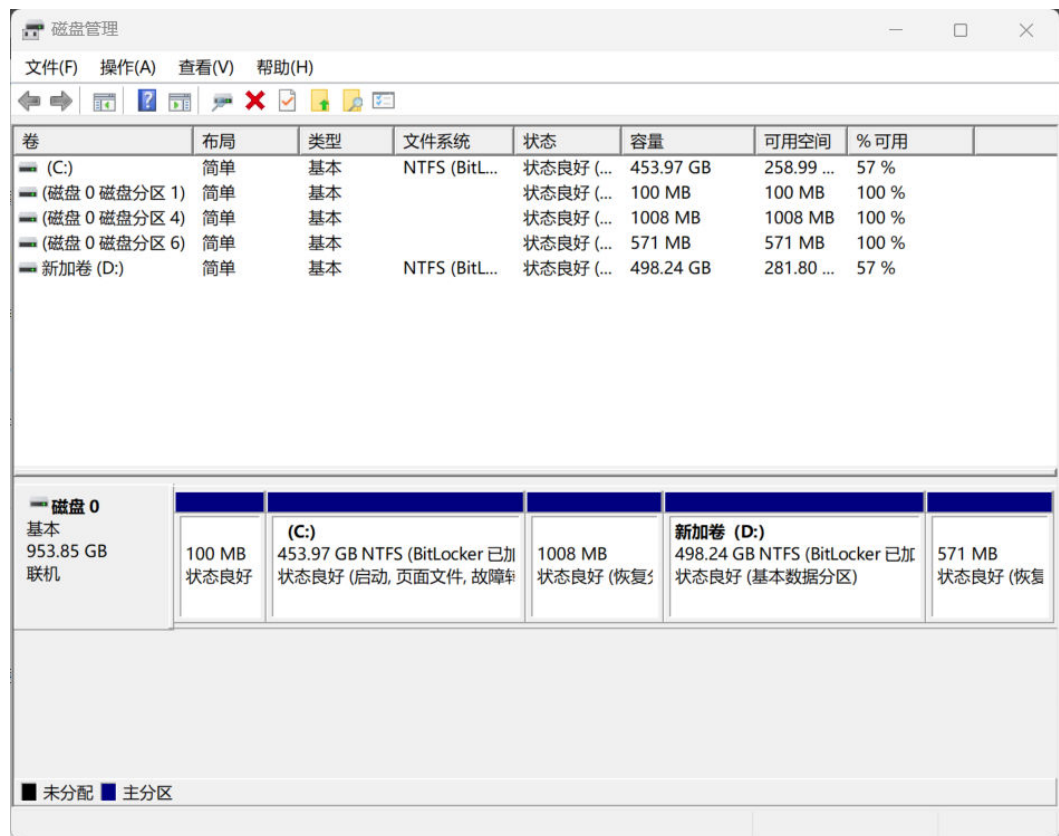
在PC已安装Windows系统的情况下，可以通过Windows的磁盘管理工具，在空闲磁盘（注意：必须为完全没有文件的空闲分区）压缩出一段空闲磁盘空间用于安装Ubuntu 22.04系统。

图 3-30 磁盘管理工具



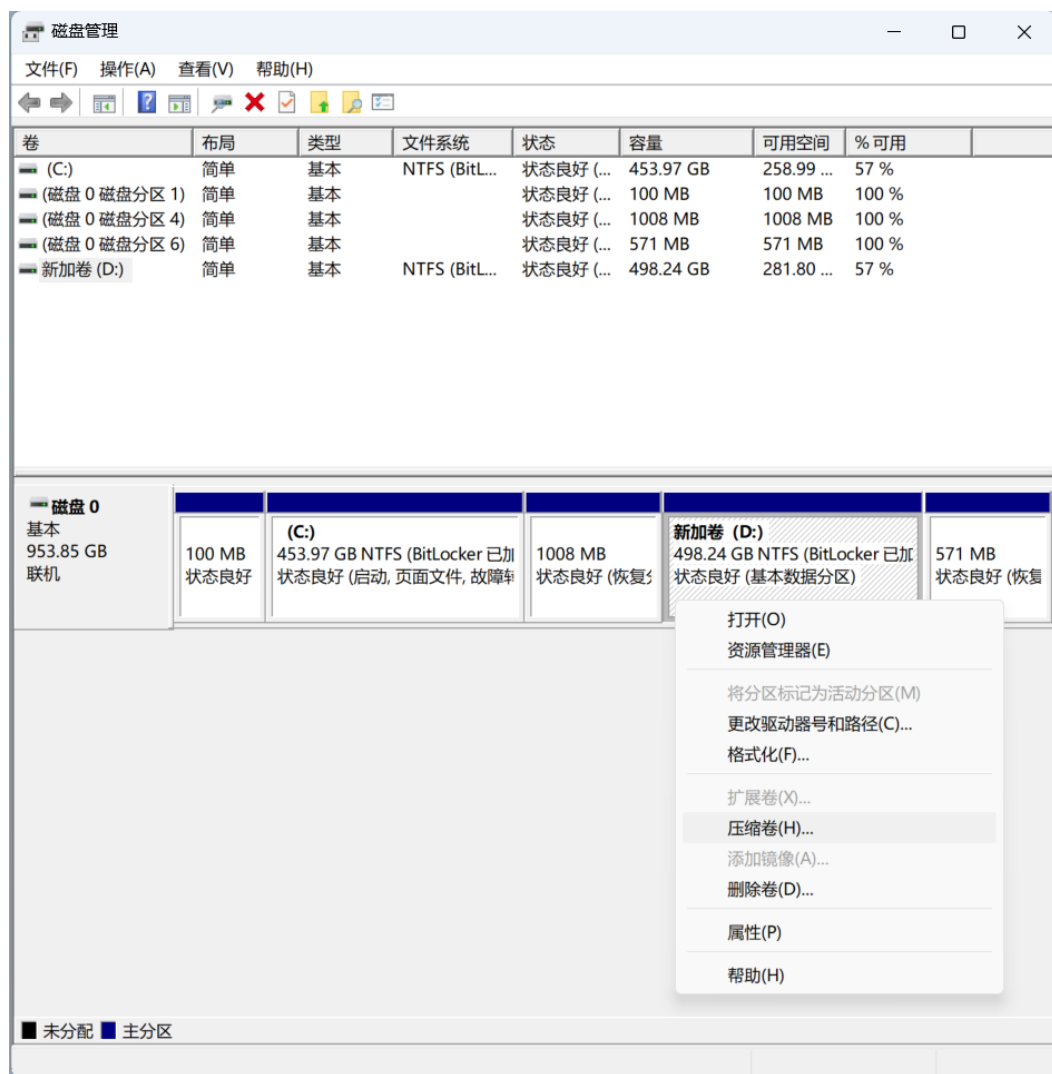
磁盘管理的界面如图3-31所示：

图 3-31 磁盘管理界面



以D盘为例，在新加卷（D:）盘上，鼠标右键，选择“压缩卷(H)”，如图3-32所示：

图 3-32 压缩卷



具体压缩大小根据用户的磁盘空闲空间大小决定，推荐>40GB，压缩出100G（102400MB）的空间，单击“压缩”按钮完成压缩。压缩后得到100GB的空闲磁盘空间用于安装Ubuntu 22.04系统。

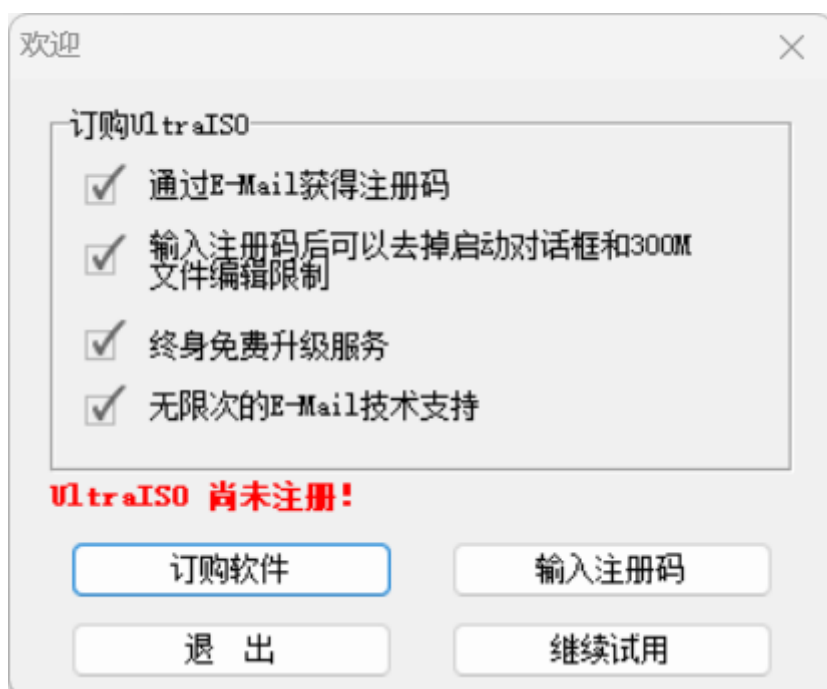
图 3-33 设置压缩大小



步骤2 制作USB系统启动盘。

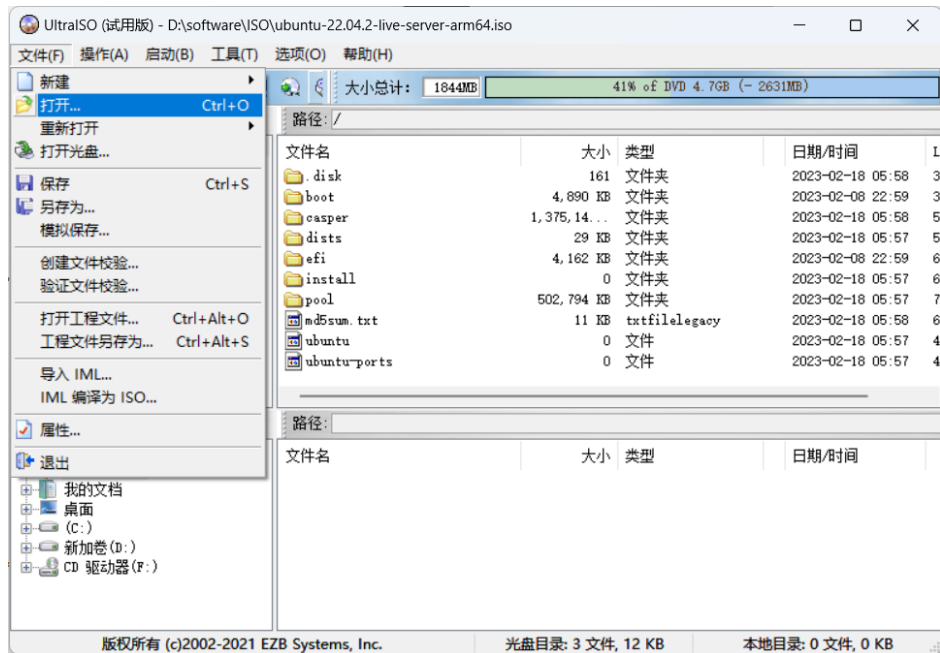
1. 准备一台Windows系统的PC, 一个空的U盘, 插入电脑的USB接口。
2. 从[Ubuntu官网](#)下载镜像文件, 推荐使用桌面版Ubuntu 22.04.2 LTS。
3. 从UltraISO官网下载[UltraISO Premium试用版本](#)进行安装。
4. 打开UltraISO, 单击“**继续试用**”。

图 3-34 UltraISO 欢迎界面



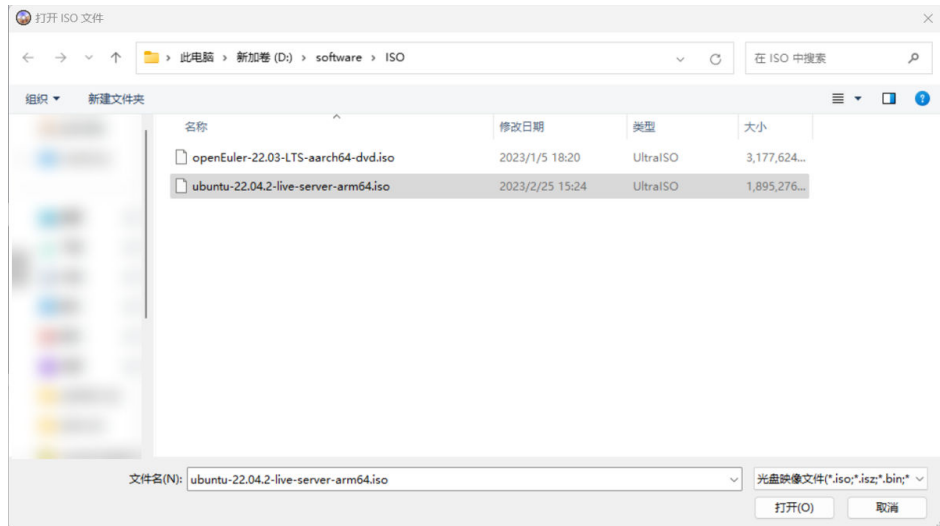
5. 在UltraISO软件界面，单击“文件”选项，单击“打开...”按钮。

图 3-35 在 UltraISO 软件界面单击打开按钮



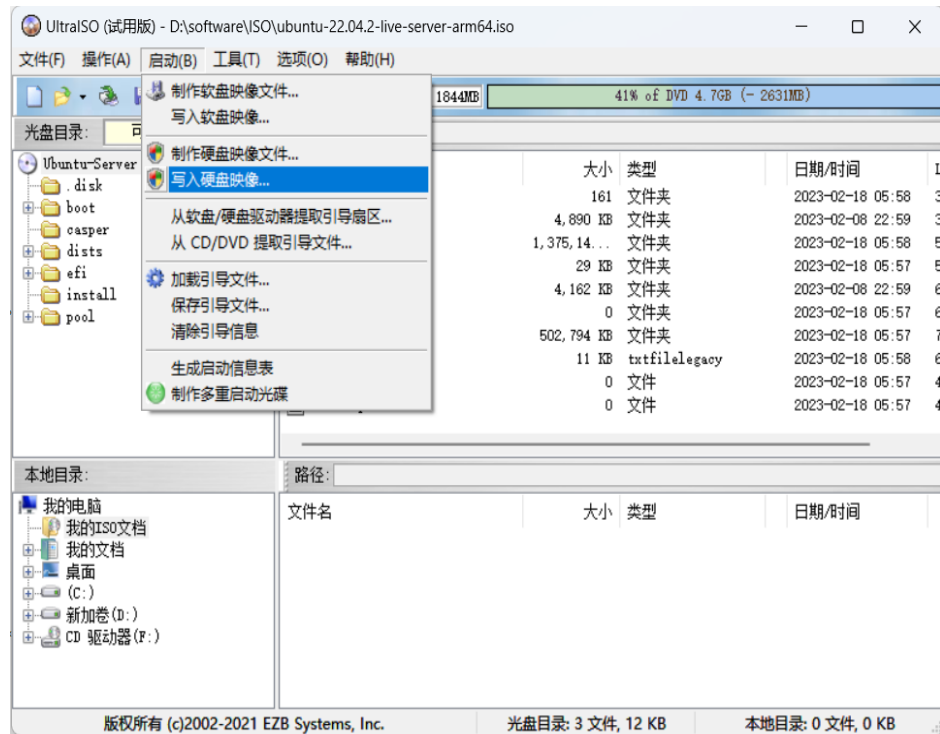
6. 打开下载的“ubuntu-22.04.2-live-server-arm64.iso”文件。

图 3-36 打开下载好的镜像文件



7. 单击“启动(B)”选项卡，单击“写入硬盘映像”按钮。

图 3-37 写入磁盘映像



8. 在弹出的界面核对U盘的盘符和容量大小是否正确。

图 3-38 核实盘符与容量大小



9. 确认无误后，单击“写入”按钮，在弹出的提示对话框中，再次核对U盘盘符和大小信息，确认无误后，单击“是(Y)”按钮，开始写入。
写入成功后，Ubuntu 22.04的系统启动盘制作完成，可以继续下面的操作。

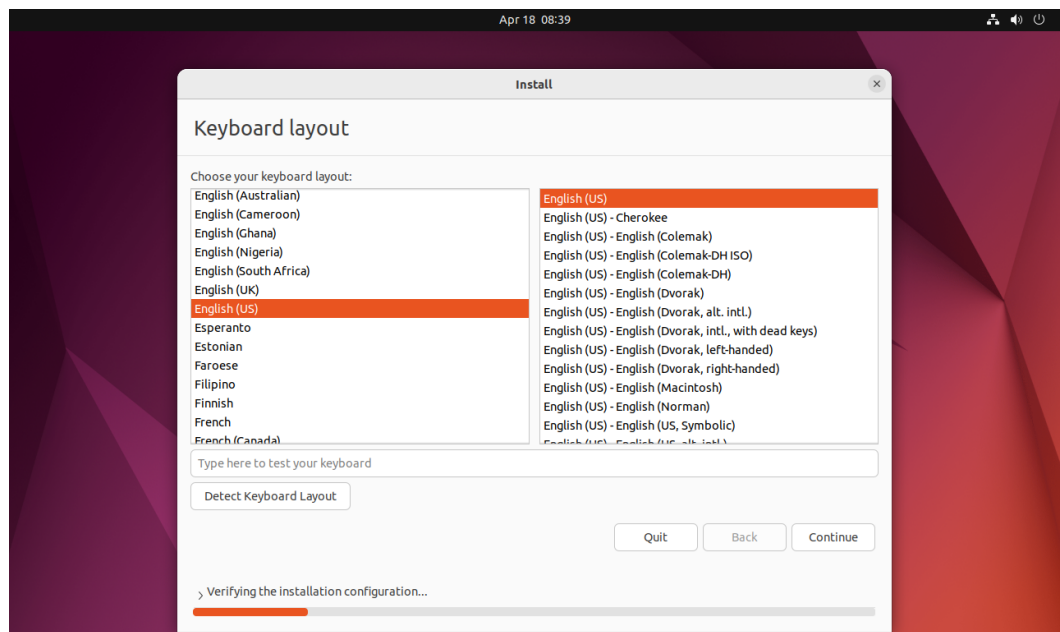
图 3-39 启动盘制作成功



步骤3 设置BIOS从USB启动

不同品牌的电脑启动进入BIOS的方式不同，大部分是在**开机启动的时候多次按del键或者F12键**即可进入BIOS，设置**打开USB启动**，然后按**F10**保存BIOS设置，保存后再次启动，选择从**U盘Ubuntu 22.04的系统启动盘启动**，即可进入Ubuntu 22.04系统的安装流程，按照提示直接安装即可。

图 3-40 安装系统界面



安装过程按照软件的提示进行安装即可，无需其他特殊配置，完成后会以创建的**普通用户**账号登录系统。

警告

在已有Windows系统的情况下，安装Windows+Ubuntu 22.04双系统，在Ubuntu 22.04系统的安装的时候，一定要选择空闲的磁盘分区来安装Ubuntu 22.04，不能选择Windows系统已经占用的磁盘分区，否则将导致Windows系统被覆盖，无法使用。

----结束

3.5.2 安装 CANN

本节介绍如何安装依赖和CANN。

步骤1 开启Windows和虚拟机之间的复制粘贴功能。

在程序菜单中找到Terminal或者使用键盘组合快捷键“Ctrl”+“Alt”+“T”打开终端，如[图3-41](#)和[图3-42](#)所示。

图 3-41 单击 LaunchPad

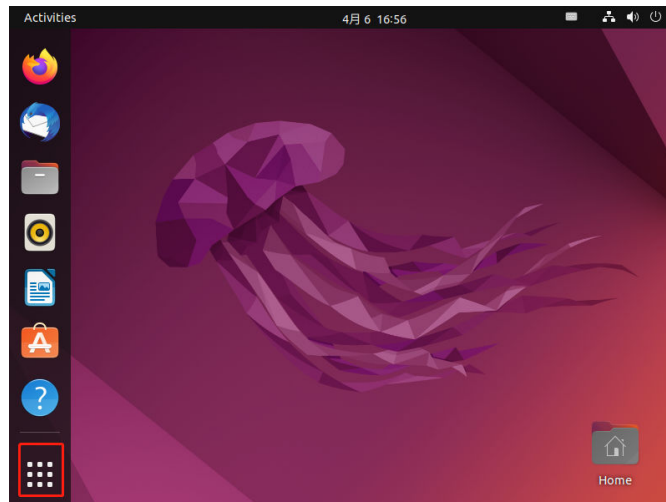
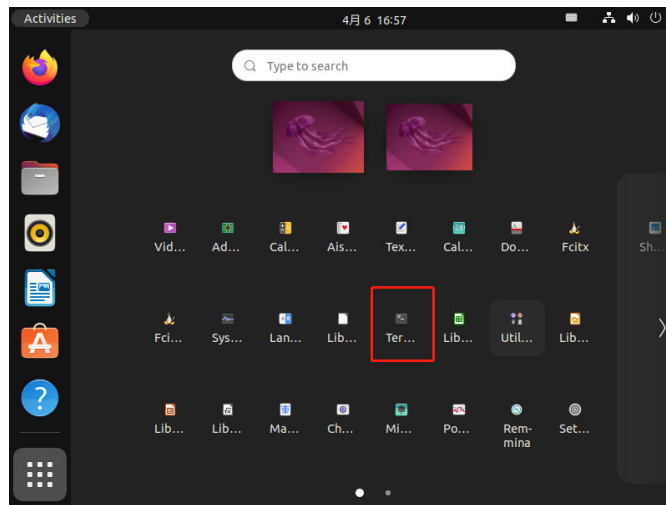


图 3-42 Terminal 工具



执行以下命令安装VMware Tools。

```
sudo apt install open-vm-tools-desktop -y
```

说明

使用sudo命令时，需输入安装Ubuntu系统时创建的用户密码。

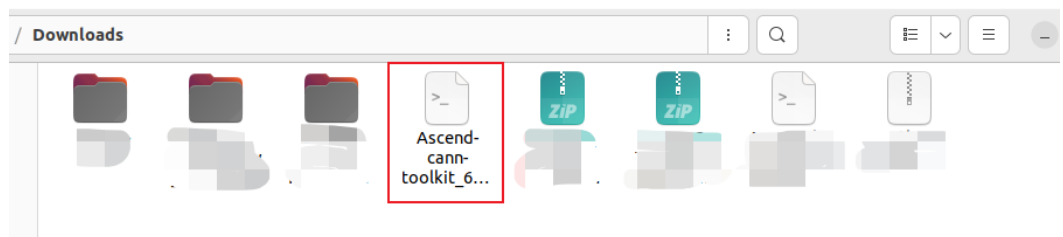
执行以下命令重启Ubuntu系统。

```
sudo reboot
```

步骤2 在Ubuntu浏览器中打开[下载链接](#)并下载CANN软件“Ascend-cann-toolkit_{version}_linux-x86_64.run”。

下载后文件会出现在“Downloads”目录中。

图 3-43 下载目录



说明

一般情况下PC为x86架构，如果现场PC为Arm架构，请下载[Arm架构安装包](#)。

步骤3 参考[安装依赖](#)安装依赖。

步骤4 执行以下命令进行开发套件包的安装。

1. 进入“Downloads”目录并打开终端，增加对软件包的可执行权限。

```
chmod +x Ascend-cann-toolkit_7.0.RC1_linux-x86_64.run
```

2. 执行以下命令安装软件。

```
./Ascend-cann-toolkit_7.0.RC1_linux-x86_64.run --install
```

安装完成后，若显示如下信息，则说明软件安装成功：

```
[INFO] xxx install success
```

xxx表示安装的实际软件包名。

步骤5 安装完成后配置开发套件包的环境变量。


```
source CANN_INSTALL_PATH/ascend-toolkit/set_env.sh
export LD_LIBRARY_PATH=CANN_INSTALL_PATH/ascend-toolkit/latest/x86_64-linux/
devlib/:$LD_LIBRARY_PATH
```

CANN_INSTALL_PATH: 为CANN软件安装目录。

----结束

3.5.3 双系统之间的文件传输

Ubuntu系统下可查看Windows系统在同一台PC的目录，且可以移动文件至任意目录。

步骤1 用户可在Ubuntu系统中点击按钮，打开文件目录。

步骤2 单击“其他位置”按钮，进入其他系统盘进行文件操作，如[图3-44](#)和[图3-45](#)所示。

图 3-44 文件目录

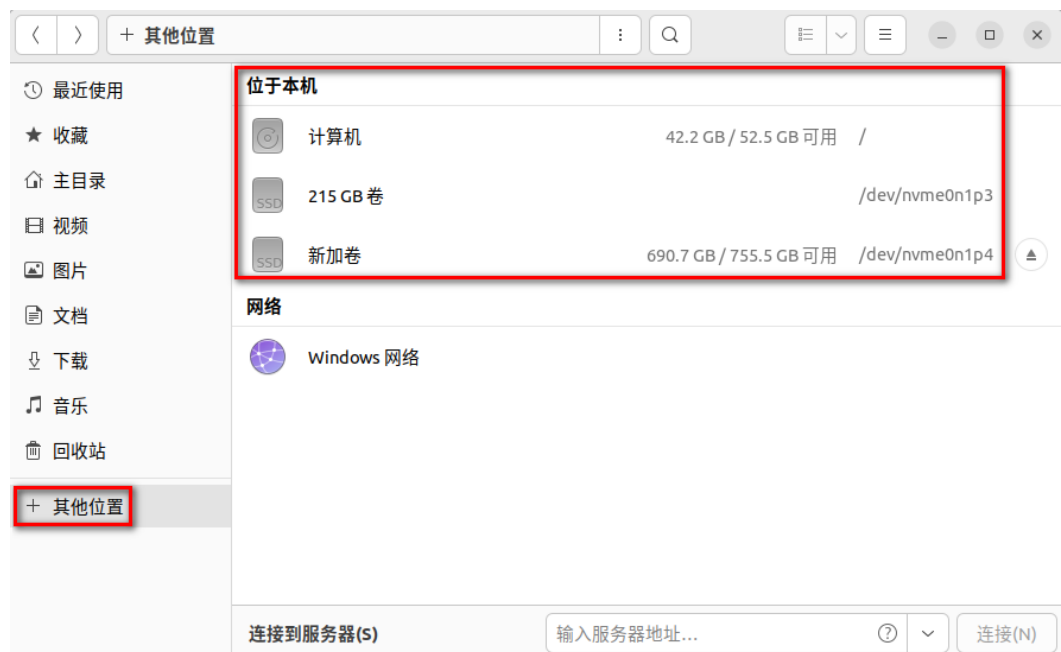
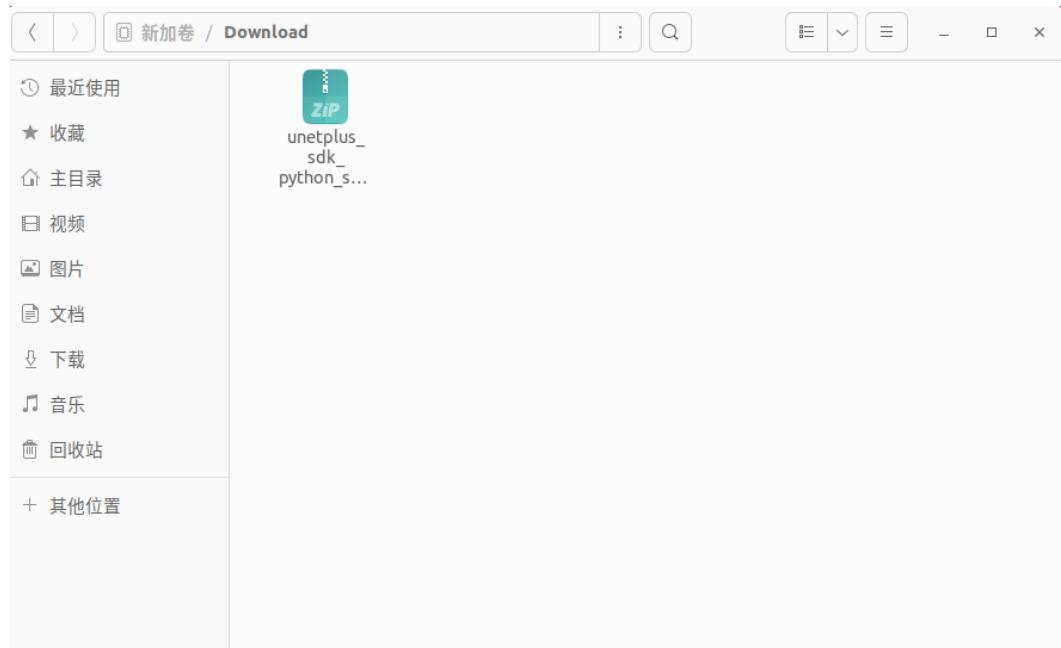


图 3-45 文件目录



----结束

3.5.4 使用 ATC 命令转换模型

本节介绍如何通过ATC工具将模型转换成支持在开发者套件上推理的离线om模型。

准备模型

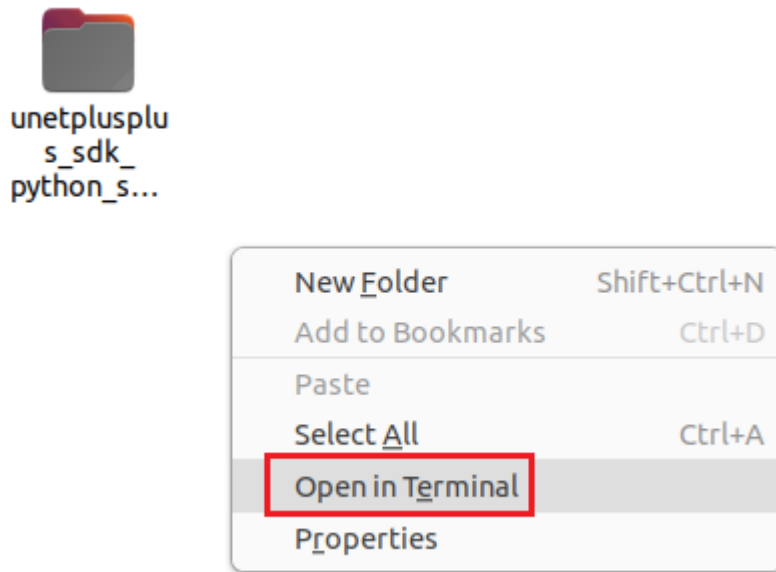
步骤1 获取网络模型。

- ONNX：单击[Link](#)或使用wget命令，解压压缩包，从“model”文件夹中获取*.onnx格式模型文件。
`wget https://ascend-repo.obs.cn-east-2.myhuaweicloud.com/Atlas%20200I%20DK%20A2/DevKit/models/sdk_cal_samples/unetplusplus_sdk_python_sample.zip`
- TensorFlow：单击[Link](#)或使用wget命令，获取*.pb格式模型文件。
`wget https://ascend-repo-modelzoo.obs.cn-east-2.myhuaweicloud.com/c-version/ResNet50_for_TensorFlow/zh/1.7/m/ResNet50_for_TensorFlow_1.7_model.zip`
- MindSpore：单击[Link](#)或使用wget命令，获取*.air格式模型文件。
`wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com:443/cannInfo/model/resnet50_export.air`
- Caffe：Caffe模型转换需要模型文件和权重文件。
 - 模型文件 (*.prototxt)：单击[Link](#)或使用wget命令下载该文件。
`wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com/003_Atc_Models/AE/ATC%20Model/resnet50/resnet50.prototxt`
 - 权重文件 (*.caffemodel)：单击[Link](#)或使用wget命令下载该文件。
`wget https://obs-9be7.obs.cn-east-2.myhuaweicloud.com/003_Atc_Models/AE/ATC%20Model/resnet50/resnet50.caffemodel`

步骤2（可选）如果是直接将模型文件下载到Ubuntu系统，则可以直接进行模型转换操作；如果是将模型文件下载到PC Windows系统，需要参考[3.5.3 双系统之间的文件传输](#)将模型上传到Ubuntu系统。

步骤3 在下载的代码文件存放目录右键单击空白处，选择“Open in Terminal”打开命令行窗口。

图 3-46 打开终端



步骤4 以下载ONNX模型为例，执行命令进入解压后模型文件存放目录。

```
cd unetplusplus_sdk_python_sample/model
```

---结束

模型转换基础示例

以下介绍模型转换必须使用的参数。

步骤1 以ONNX模型为例，执行如下命令生成离线模型（如下命令中使用的目录以及文件均为样例），模型转换必须使用的参数介绍如表3-8所示。

```
atc --model=model.onnx --framework=5 --output=model --soc_version=Ascend310B4
```

📖 说明

- 若用户在执行模型转换命令前关闭了配置环境变量的终端窗口，需参考步骤3与步骤5，在新终端窗口再次配置Python以及CANN的环境变量。
- 转换模型失败，可参见[应用进程占用内存超出限制导致系统异常终止](#)解决。

表 3-8 参数说明

参数名	参数说明
--model	原始模型文件，填写模型文件时需要带上格式，如.onnx。
--weight	原始模型权重，该参数在转换Caffe模型场景下使用，其他框架不使用。
--framework	原始框架类型，各框架对应的数值如下： 0:Caffe; 1:MindSpore; 3:Tensorflow; 5:ONNX
--output	保存转换后的om离线推理模型文件路径。

参数名	参数说明
-- soc_version	昇腾AI处理器型号，填写“Ascend310B4”。

步骤2 若提示如下信息，则说明模型转换成功，若模型转换失败，则请参见[错误码参考](#)进行定位。

```
ATC run success
```

成功执行命令后，在--output参数指定的路径下，可查看离线模型（如：model.om）。

模型编译时，若遇到AI CPU算子不支持某种数据类型导致编译失败的场景，可通过启用Cast算子自动插入特性快速将输入转换为算子支持的数据类型，从而实现网络的快速打通，详细流程请参见[开启AI CPU Cast算子自动插入特性](#)。

----结束

其他常用转换参数说明

[模型转换基础示例](#)介绍了模型转换必须使用的参数，接下来以YoloV5模型和SVTR模型为例介绍其他常用参数的使用方法。更多模型转换参数说明请参见《[使用ATC工具转换模型](#)》中“参数说明”章节。

单击链接或使用wget命令下载[YoloV5模型](#)代码包，在model目录中获取onnx模型文件。

```
wget https://ascend-devkit-tool.obs.cn-south-1.myhuaweicloud.com/models/yolo_sdk_python_sample.zip
```

YoloV5模型转换命令如下：

```
atc --model=yolov5s.onnx --framework=5 --output=yolov5s_bs1 --input_format=NCHW --  
soc_version=Ascend310B4 --input_fp16_nodes="images"
```

单击链接或使用wget命令下[SVTR模型](#)代码包，在models目录中获取onnx模型文件。

```
wget https://ascend-devkit-tool.obs.cn-south-1.myhuaweicloud.com/models/ocr_acl_sample.zip
```

SVTR模型转换命令如下：

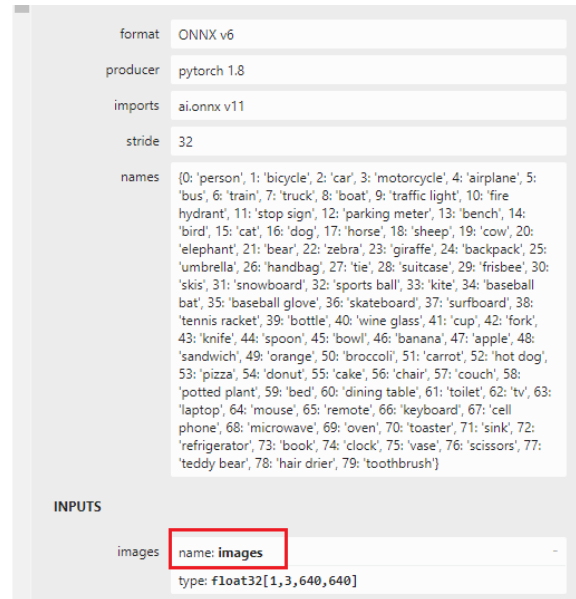
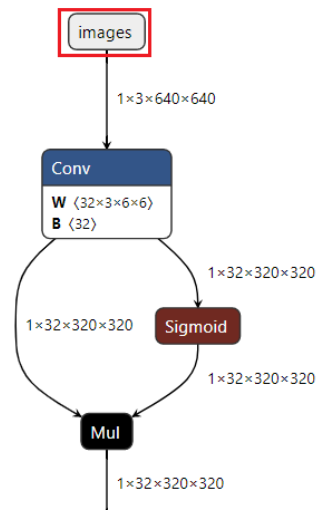
```
atc --model=svtr.onnx --framework=5 --input_shape='x:1,3,48,1440' --input_format=NCHW --  
soc_version=Ascend310B4 --output=svtr
```

表 3-9 参数说明

参数名	参数说明
--input_format	<p>输入Tensor的内存排列方式，NCHW指代batch、channels、height、width。</p> <ul style="list-style-type: none">当原始框架为Caffe时，支持NCHW、ND（表示支持任意维度格式，$N \leq 4$）两种格式，默认为NCHW。当原始框架为ONNX时，支持NCHW、NCDHW、ND（表示支持任意维度格式，$N \leq 4$）三种格式，默认为NCHW。当原始框架是TensorFlow时，支持NCHW、NHWC、ND、NCDHW、NDHWC五种输入格式，默认为NHWC。<ul style="list-style-type: none">如果TensorFlow模型是通过ONNX模型转换工具输出的，则该参数必填，且值为NCHW。如果原始模型中含有带data_format入参的算子，则该参数必填，推荐取值为ND，模型转换过程中会根据data_format属性的算子，推导出具体的format。若用户无法确定输入数据格式，则推荐指定为ND。当原始框架为MindSpore时，只支持配置为NCHW。 <p>一般情况下不需要使用该参数，如果用户开发的应用代码前处理对内存排列有要求，可以使用该参数并填写所需的内存排列方式。</p>
--input_shape	<p>模型的输入节点名称和shape，shape的格式一般为[batch,channels,height,width]。</p> <p>一般情况下不需要使用该参数，如果要转换的模型为动态shape的ONNX模型时，需要使用该参数并填写shape。</p> <p>本文以将一个动态shape的SVTR模型转换为静态om模型为例。</p>
--input_fp16_nodes	<p>指定输入数据类型为FP16的输入节点名称。若不指定，则默认是float32数据类型。</p> <p>此参数可根据用户需要选择是否指定，若为默认float32，则精度相对较高；若为float16，则性能相对较高，在精度无明显下降的情况下有利于性能的提升。</p>

- 输入节点名称查看方法
使用[Netron模型可视化工具](#)打开（单击Open Model按钮）PC本地的模型文件，单击输入节点，查看输入节点名称。

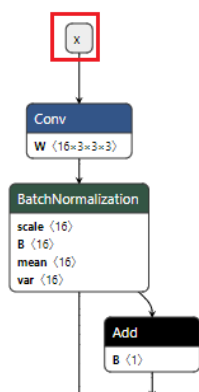
图 3-47 查看输入节点名称



- 输入节点shape查看方法和input_shape参数设置策略

使用Netron模型可视化工具打开模型，单击输入节点查看shape，可以看出输入节点x的shape为[p2o.DynamicDimension.0,3,48,p2o.DynamicDimension.1]，可以看出这个模型的输入是NCHW格式。第一个输入和最后一个输入分别为batch_size和width，且他们都由一个string填充，这种情况该维度为动态参数值（有时batch_size和width值为-1，这种情况和由string填充等价）。此时如果将该动态shape模型转换为batch size为1，宽度为1440的静态om模型时，input_shape参数可以设置为[1,3,48,1440]，1440为shape中的width，可以按需设置。

图 3-48 查看 shape



3.6 FAQ

3.6.1 运行 wsl --status 报错

现象描述

运行wsl --status查看wsl状态时报虚拟化错误，如图3-49所示。

图 3-49 虚拟化报错

```
默认版本：2
当前计算机配置不支持 WSL2。
请启用“虚拟机平台”可选组件，并确保在 BIOS 中启用了虚拟化。
有关信息，请访问https://aka.ms/enablevirtualization
```

可能原因

运行虚拟化软件或者其他应用场景时，禁用了Hyper-V。

解决方法

解决Windows 10、Windows 11系统运行wsl --status查看wsl状态时报虚拟化错误，可以通过设置->应用->相关设置->启用或关闭Windows功能->勾选Hyper-V来进行设置，详情如下所示。

图 3-50 Windows 设置及应用界面

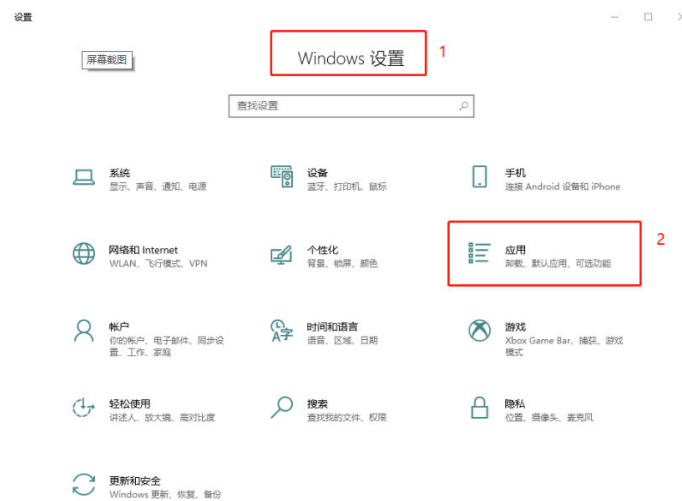


图 3-51 相关设置界面

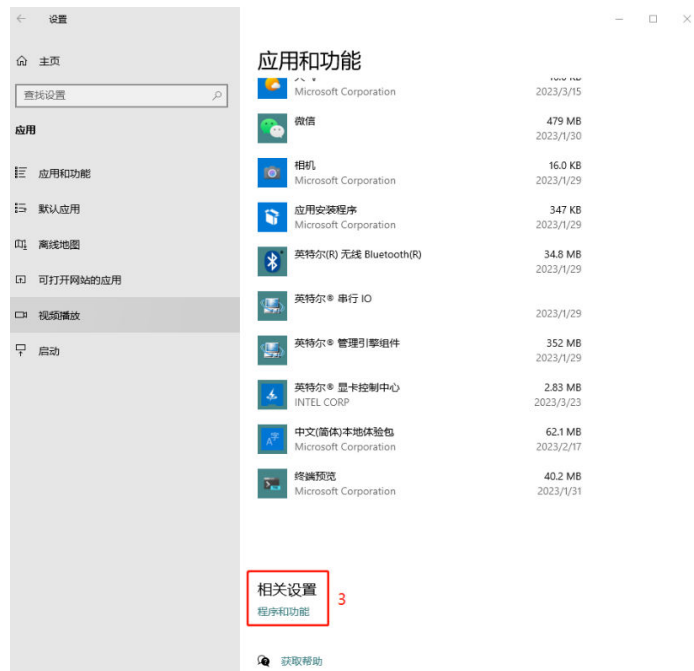


图 3-52 启用或关闭 Windows 功能界面

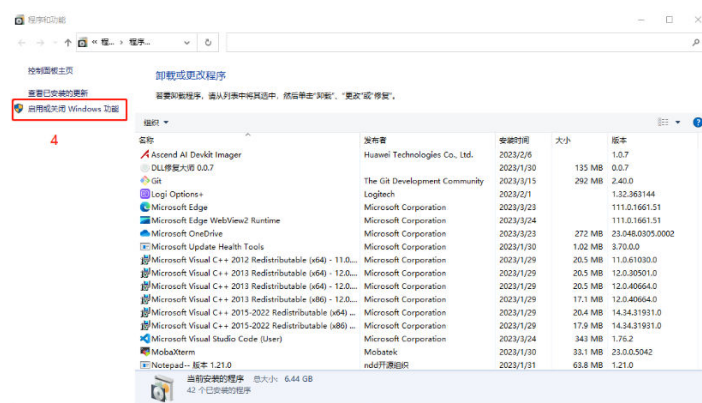
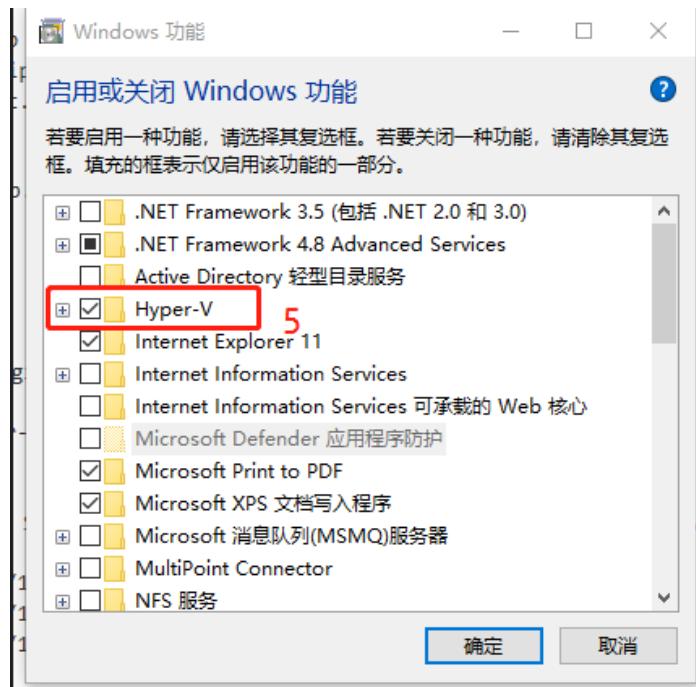


图 3-53 启用或关闭 Windows 功能界面并勾选 Hyper-V



3.6.2 运行 wsl --update 报错

现象描述

运行wsl --update时报错码为0x80240438。

可能原因

通常是由于网络问题导致。

- 请检查当前网络的连接状态以及是否能够访问Internet。
- 检查代理服务器。
- 检查防火墙是否阻止了wsl的网络连接。

若网络连接正常，但仍出现相同错误码，请自行搜索解决方案。

3.6.3 无法解析服务器的名称或地址

现象描述

```
PS C:\Windows\system32> wsl --list --online  
无法解析服务器的名称或地址
```

解决方法

可能与DNS、网络有关。

家庭网络无法访问<http://githubusercontent.com>，尝试VPN或修改Hosts文件。

3.6.4 在 PC 安装 Linux Ubuntu 22.04 后无法启动

现象描述

在安装Ubuntu操作系统时，提示This computer uses Intel RST (Rapid Storage Technology) .You need to turn off RST before installing Ubuntu.信息。

可能原因

当前BIOS设置，开启了英特尔快速存储技术（RST）设置，与Ubuntu系统不兼容运行。

解决方法

- 步骤1** 重新启动电脑，并进入BIOS设置。
- 步骤2** 在配置设置中关闭将存储模式从“RAID mode”切换为“AHCI mode”。
- 步骤3** 保存设置并重启电脑，再次进入Ubuntu系统安装界面进行安装。

----结束