**一、端边云产业赋能系统大模型迁移技术路线**

大模型虽然性能高，但因其计算量参数量计算量大难以在实际中部署使用，因此有必要将其所学知识迁移到小模型上，再将小模型部署使用。

**1. 大模型迁移框架技术路线**

将大模型知识迁移到小模型，其中较为成熟的技术是知识蒸馏，其通过在下游任务的迁移过程中对齐大模型和小模型的特征以实现知识从大模型到小模型的蒸馏。如此蒸馏迁移得到的小模型通常要比直接在下游任务上单独训练小模型的性能要更好。

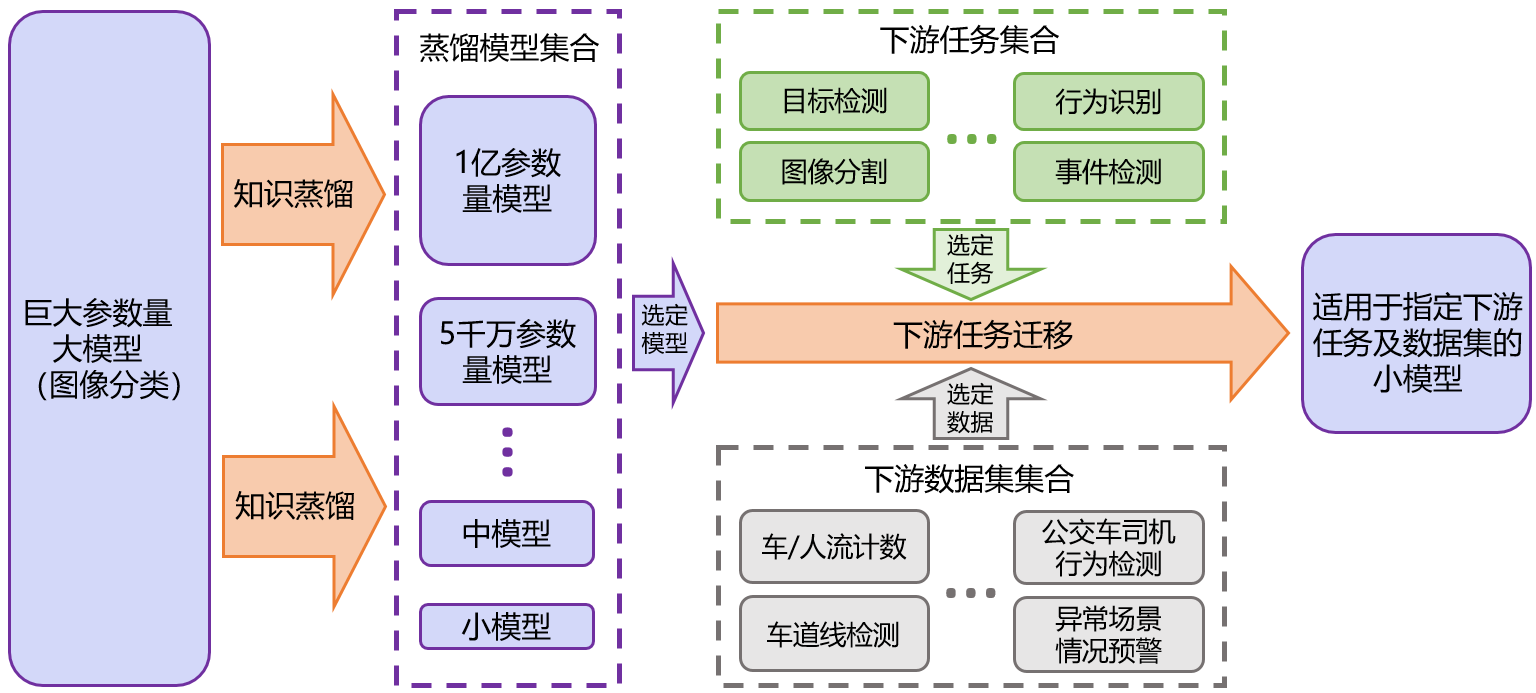


图1大模型蒸馏迁移至下游任务的技术路线示意图

大模型蒸馏迁移的框架图如图1所示，主要流程如下：

（1）给定一个大模型，选定一个小模型(参数量从1千万到1亿不等) ；

（2）选定下游任务（如目标检测、图像分割）和下游数据集；

（3）使用选定的数据集，在选定的任务框架上，替换骨干网络为选定的小模型并进行蒸馏训练，最后整个流程结束之后输出适用于指定下游任务的小模型。

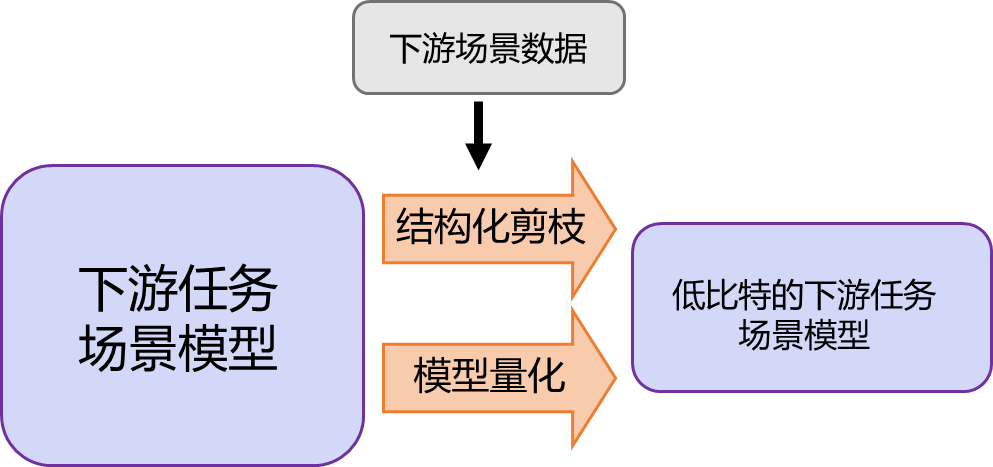


图2场景模型的模型压缩技术路线示意图

知识蒸馏往往只能对选定的小模型进行蒸馏，如果该小模型在实际中仍然参数量计算量过大，那么则需要进一步借助模型压缩技术对小模型进行压缩，如图2所示。值得一提的是，在压缩过程中仍需要场景数据支持。

（1）结构化剪枝：去除冗余通道，减少模型参数量与计算量。

（2）模型量化：使用低比特整数表示网络权重与激活值，加速硬件推理，降低功耗。

**2. 场景模型转换技术路线**

在小模型迁移训练完成之后，其还需转换到推理硬件（如ARM设备，FPGA芯片）和推理框架（如MindSpore Lite）支持的格式。这一过程如图3所示，以下游场景模型的参数和计算图为输入，将其移植转换到CPU端、GPU端、ARM端、鹏城自研芯片上，提供移植转换后的模型文件。

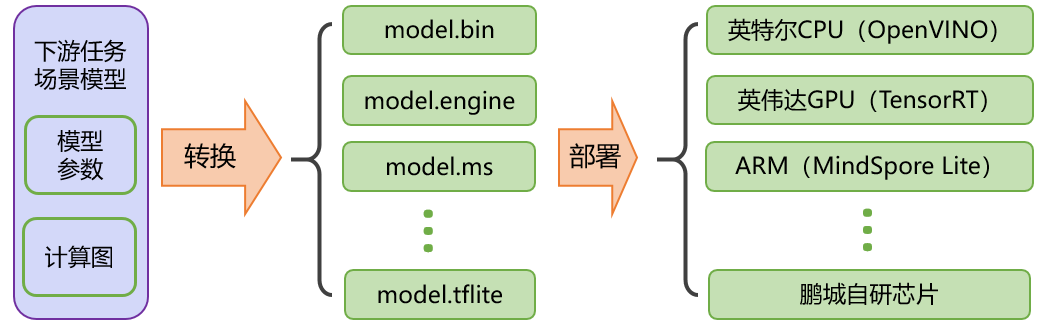


图3 模型转换技术路线示意图

具体地，首先加载场景模型的模型定义代码文件和.ckpt参数文件，保存为同时带有计算图和模型参数的.onnx格式的模型文件或.mindir格式的模型文件。然后，对于多种设备类型的转换细节如下。

1. 面向英特尔CPU环境，加载上一步得到的.onnx格式的模型文件，转换为可在OpenVINO推理框架上运行的.bin格式的模型文件。
2. 面向英伟达GPU环境，加载上一步得到的.onnx格式的模型文件，转换为可在TensorRT推理框架上运行的.engine格式的模型文件。
3. 面向ARM端环境，加载上一步得到的.mindir格式的模型文件，转换为可在MindSpore Lite推理框架上运行的.ms格式的模型文件。
4. 面向鹏城自研芯片环境，加载上一步得到的.onnx格式的模型文件，转换为可在自研芯片上运行的.tflite格式的模型文件。