【教程】在AMD GPUs平台部署和运行Olive优化版Automatic1111 Stable Diffusion WebUI

[更新]: Automatic1111（后简称A1111）WebUI界面目前已经支持Microsoft Olive的DirectML分支，它能生成优化模型并在A1111 WebUI下运行，无需单独针对AMD平台进行优化。

你是否知道，在Windows平台使用Microsoft DirectML，我们可以大幅提高Olive版Stable Diffusion的生成速度？Microsoft和AMD持续在软硬件方面进行合作，通过优化Microsoft DirectML API和AMD用户模式驱动程序机器学习的软件层，使用户能够获得基于AMD GPUs的更强大的AI能力。

A graph on a white background

Description automatically generated

#### Microsoft Olive概述

Olive是一款对模型进行转换、优化、量化和自动调整的Python工具，通过例如DirectML等ONNX Runtime获得最佳的AI推理性能。Olive通过提供线性的优化工具链来整合优化技术，极大的简化了模型处理，这对于像Stable Diffusion这样的复杂模型尤为重要。用于Stable Diffusion的DirectML包含了以下技术：

* 模型转换：

将基本模型从PyTorch转换为ONNX

* Transformer模型图形优化：

将子图形融合到multi-head attention算子中，改善低效率的转换过程

* 量化：

将大量FP32转换为FP16层，减少模型的VRAM占用并提高性能

因此，上述优化使DirectML显著提高AMD GPUs处理Transformer模型推理任务，例如Stable Diffusion的推理效率。

#### 安装前提条件

* 安装Git for Windows

https://git-scm.com/downloads

* 安装Anaconda

https://www.anaconda.com/download/

* 具备AMD独立显卡的平台

https://www.amd.com/zh-hans/graphics/radeon-rx-graphics

* 驱动程序AMD Software: Adrenalin Edition 23.7.2或更高版本

https://www.amd.com/zh-hans/support

#### 在AMD GPUs上使用Olive优化版A1111 Stable Diffusion WebUI

打开Anaconda终端，输入以下命令，然后按回车键：

* 创建名为Automatic1111\_olive的conda虚拟环境

conda create --name Automatic1111\_olive python=3.10.6

* 启动虚拟环境

conda activate Automatic1111\_olive

* 克隆stable diffusion webui到本地

git clone https://github.com/lshqqytiger/stable-diffusion-webui-directml

* 进入文件夹

cd stable-diffusion-webui-directml

* 初始化子模型

git submodule update --init –recursive

* 启动A1111 WebUI

webui.bat --onnx --backend directml

这一步将会安装Olive、ONNX Runtime以及其他软件依赖项并自动启动WebUI，这个过程可能需要几分钟

在弹出的WebUI窗口中，转到Olive优化选项卡并点击优化按钮

A screenshot of a computer

Description automatically generated

选择如图所示的模型

A white rectangle with black lines

Description automatically generated

进入’txt2img’开始生成第一张图片吧

A screenshot of a computer

Description automatically generated

运行在默认PyTorch路径下，AMD Radeon RX 7900 XTX的推理能力为1.22 iterations/second

运行在Microsoft Olive优化模型下，AMD Radeon RX 7900 XTX的推理能力为20.47 iterations/second

AMD Radeon RX 7900 XTX的推理能力提升高达15.8倍