赛题一：Softmax 算子

1. 赛题背景

Softmax 是深度学习模型中最基础且高频使用的激活函数之一，广泛应用于多分类任务及 Transformer 架构（如 Attention 机制）中。在在大模型（LLM）长序列推理场景下，Softmax 的显存读写带宽往往成为性能瓶颈。

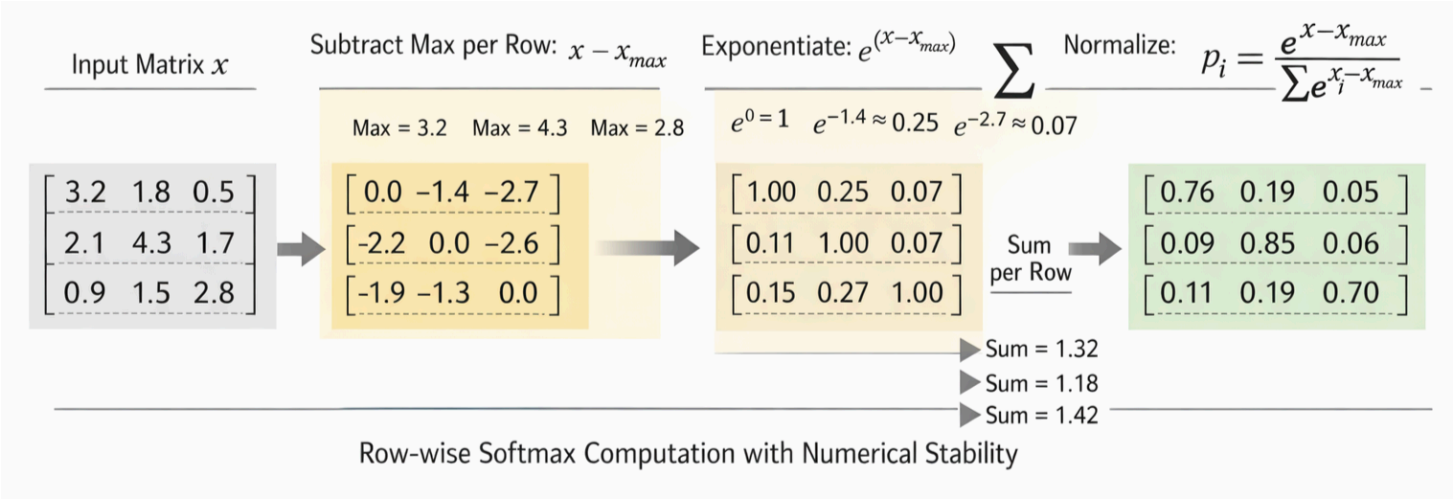
本赛题要求选手在 **Ascend 910B3** 硬件平台上，使用 **Triton-Ascend** 或 **TileLang-Ascend** 编写 Softmax 算子。

2. 算子功能

一个简单的 torch 实现如下：

```
def naive_softmax(x):
    # 求出每行的最大值
    x_max = x.max(dim=1)[0]
    # 每个元素减去对应行的最大值
    z = x - x_max[:, None]
    # 对每个元素作 exp 指数运算
    numerator = torch.exp(z)
    # 以下代码对每行元素作归一化处理，使得每行元素和等于 1
    denominator = numerator.sum(dim=1)
    ret = numerator / denominator[:, None]
    return ret
```

一个简单的示意图如下：



3. 输入输出

输入/输出张量元素类型均为 `float32` ，形状均为 [512, 256]。Softmax 在最后一个维度上进行规约。

4. 答题要求

代码框架在平台的共享文件存储中已经给出（权限为只读），路径如下。

存储类型： 文件存储 源路径： /softmax 挂载路径： /1-softmax/softmax

请你在已给出的 `softmax_triton.py` 和 `softmax_tilelang.py` 中任选其一，补全其中的 **todo**，并运行文件通过结果正确性测试。请在希冀平台的对应作业中提交你的**代码文件**和**运行结果截图**。

注：你可以适当的合理修改给出的框架中 `softmax_kernel` 和 `softmax` 函数的内容。

5. 拓展

本次题目并不强制要求性能，结果正确即可。开动你聪明的脑筋想想，怎么在实现 Softmax 算子时提高算子性能呢？（不作强制要求）