# 自定义 CUDA 函数的轻量级包装器 —— bitsandbytes篇

来自： AiGC面试宝典

宁静致远

2024年01月27日 19:14





## 一、什么是 bitsandbytes?

bitsandbytes 是自定义 CUDA 函数的轻量级包装器，特别是 8 比特优化器、矩阵乘法和量化函数。 主要特征如下：

- 具有混合精度分解的 8 比特矩阵乘法
- LLM.int8() 推理
- 8 比特优化器：Adam、AdamW、RMSProp、LARS、LAMB、Lion（节省 75% 内存）
- 稳定的嵌入层：通过更好的初始化和标准化提高稳定性
- 8 比特量化：分位数、线性和动态量化
- 快速的分位数估计：比其他算法快 100 倍

## 二、如何才能使用 bitsandbytes?

量化模型的唯一条件是包含 torch.nn.Linear 层，因此量化对于任何模态都可以实现开箱即用。用户可以开箱即用地加载诸如 Whisper、ViT、Blip2 之类的 8 比特或 4 比特(FP4/NF4)模型。

## 三、如何使用 bitsandbytes?

使用 NF4 量化加载 4 比特模型的示例：

```
from transformers import BitsAndBytesConfig
nf4_config = BitsAndBytesConfig(
   load_in_4bit=True,
   bnb_4bit_quant_type="nf4",
   bnb_4bit_use_double_quant=True,
   bnb_4bit_compute_dtype=torch.bfloat16
)

model_nf4 = AutoModelForCausalLM.from_pretrained(model_id,
quantization_config=nf4_config)
```

使用 FP4 量化加载 4 比特模型的示例：

```
import torch
from transformers import BitsAndBytesConfig
```

```
quantization_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_compute_dtype=torch.bfloat16
)
```