# Mamba 的 CUDA 实现解读

#### 黄有

## 1 Mamba 模型架构

本文只针对 Mamba [1] 核心部分进行解读,Mamba 整体架构可以参考:一文读懂 Mamba,以及 Mamba 所采用的离散化方式: 状态空间模型 SSM 的离散化过程推导。基于此,本文直接采用最终的离散化形式进行后续的代码解读,具体如下

$$egin{aligned} oldsymbol{h}_k &= e^{\Delta A} oldsymbol{h}_{k-1} + \left( e^{\Delta A} - I 
ight) A^{-1} B oldsymbol{x}_k \ oldsymbol{y}_k &= C oldsymbol{h}_k + D oldsymbol{x}_k, \end{aligned}$$

其中  $x_k, y_k$  可以理解为当前网络层在单个 Token(对于 NLP 任务)上的输入和输出,k 指定了 Token 的索引, $h_k$  是对应于这个 Token 的隐藏状态 (特征),其他变量都是参数(可以固定,也可以随着输入  $x_k$  而变化)。为简化后续内容,设置维度:

$$egin{aligned} oldsymbol{x}_k, oldsymbol{y}_k \in \mathbb{R}^d, \ oldsymbol{h}_k \in \mathbb{R}^n, \ oldsymbol{\Delta} \in \mathbb{R}, \ oldsymbol{A} \in \mathbb{R}^{n imes n}, \ oldsymbol{B} \in \mathbb{R}^{n imes d}, \ oldsymbol{C} \in \mathbb{R}^{d imes n}, \ oldsymbol{D} \in \mathbb{R}^{d imes d}, \end{aligned}$$

以上是 Mamba 的理论结构,实际 Mamba 的实现形式略有变化,以下结合代码进行讲解。

#### 2 Mamba 代码结构

以下是Mamba 源码 (更新截止 2024 年 2 月 24 日)的目录结构:

```
1 mamba/
2 -- csrc
      L— selective_scan
4
           -- reverse_scan.cuh
           -- selective_scan_bwd_bf16_complex.cu
           -- selective_scan_bwd_bf16_real.cu
6
           -- selective_scan_bwd_fp16_complex.cu
7
           -- selective_scan_bwd_fp16_real.cu
9
           -- selective_scan_bwd_fp32_complex.cu
10
           -- selective_scan_bwd_fp32_real.cu
11
           -- selective_scan_bwd_kernel.cuh
           -- selective_scan_common.h
12
           -- selective_scan.cpp
13
          -- selective_scan_fwd_bf16.cu
14
          -- selective_scan_fwd_fp16.cu
          -- selective_scan_fwd_fp32.cu
16
           -- selective_scan_fwd_kernel_comment.cuh
18
           -- selective_scan_fwd_kernel.cuh
           --- selective_scan.h
19
           --- static_switch.h
20
           L— uninitialized_copy.cuh
21
22 --- mamba_ssm
       -- __init__.py
23
       --- models
           -- config_mamba.py
25
26
            -- __init__.py
           L— mixer_seq_simple.py
        -- modules
28
           ├-- __init__.py
29
           L—— mamba_simple.py
30
        - ops
           -- __init__.py
32
33
           -- selective_scan_interface.py
           L-- triton
34
               -- __init__.py
35
               - layernorm.py
               L—— selective_state_update.py
37
         utils
           - generation.py
39
40
            -- hf.py
           L__ __init___.py
41
42
```

参考文献 3

以上罗列了源码中和 Mamba 模型相关部分的目录结构,更具体的代码结构可以参考下面的思维导图:

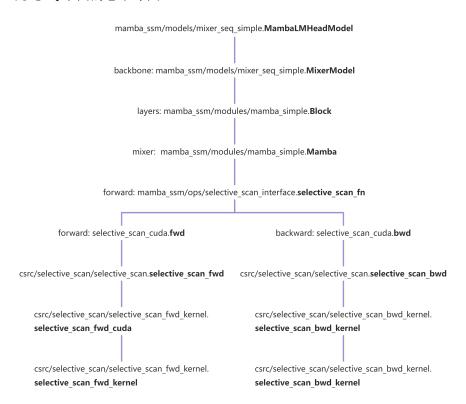


图 1: Mamba 代码结构

### 参考文献

[1] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023.