

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

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

其中 $\mathbf{x}_k, \mathbf{y}_k$ 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， \mathbf{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 \mathbf{x}_k 而变化）。为简化后续内容，设置维度：

$$\begin{aligned}\mathbf{x}_k, \mathbf{y}_k &\in \mathbb{R}^d, \\ \mathbf{h}_k &\in \mathbb{R}^n, \\ \Delta &\in \mathbb{R}, \\ A &\in \mathbb{R}^{n \times n}, \\ B &\in \mathbb{R}^{n \times d}, \\ C &\in \mathbb{R}^{d \times n}, \\ D &\in \mathbb{R}^{d \times d},\end{aligned}$$

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
3 |   |— selective_scan
4 |       |— reverse_scan.cuh
5 |       |— selective_scan_bwd_bf16_complex.cu
6 |       |— selective_scan_bwd_bf16_real.cu
7 |       |— selective_scan_bwd_fp16_complex.cu
8 |       |— 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
12 |      |— selective_scan_common.h
13 |      |— selective_scan.cpp
14 |      |— selective_scan_fwd_bf16.cu
15 |      |— selective_scan_fwd_fp16.cu
16 |      |— selective_scan_fwd_fp32.cu
17 |      |— selective_scan_fwd_kernel_comment.cuh
18 |      |— selective_scan_fwd_kernel.cuh
19 |      |— selective_scan.h
20 |      |— static_switch.h
21 |      |— uninitialized_copy.cuh
22 |— mamba_ssm
23 |   |— __init__.py
24 |   |— models
25 |       |— config_mamba.py
26 |       |— __init__.py
27 |       |— mixer_seq_simple.py
28 |   |— modules
29 |       |— __init__.py
30 |       |— mamba_simple.py
31 |   |— ops
32 |       |— __init__.py
33 |       |— selective_scan_interface.py
34 |       |— triton
35 |           |— __init__.py
36 |           |— layernorm.py
37 |           |— selective_state_update.py
38 |   |— utils
39 |       |— generation.py
40 |       |— hf.py
41 |       |— __init__.py
42 |— ...

```

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.