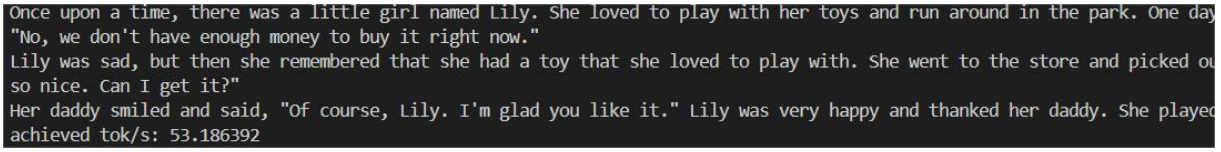
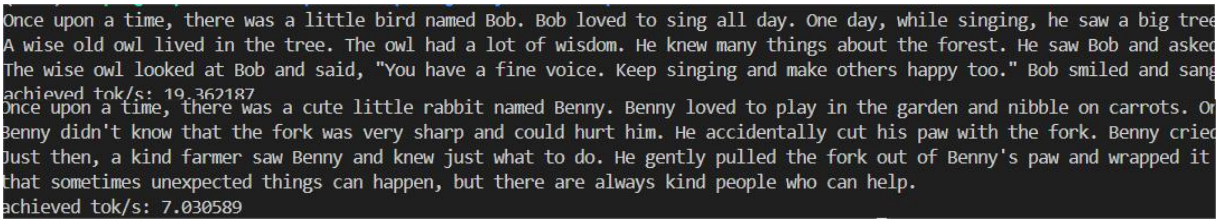
计算机体系结构实验 6 实验报告

实验流程

运行 **Llama2** 模型

截图如下所示





打开终端， 观察结果，显然模型参数越大， `text` 生成越慢。

优化矩阵乘法

修改代码进行优化

采用CUDA进行优化在main函数中给每次需要矩阵乘法的三个矩阵分配显存（直接用最大值分配）

```
int main(int argc, char *argv[]) {

     cudaMalloc((void**)&d_x, 2048 *sizeof(float));
     cudaMalloc((void**)&d_w, 2048*32000 * sizeof(float));
     cudaMalloc((void**)&d_xout, 32000 * sizeof(float));
    ...
    }
```

```
__global__void matmulKernel(float* xout, const float* x, const float* w, int n,
int d) {
    // 每个线程负责计算 xout 的一个元素
    int i = blockIdx.x * blockDim.x + threadIdx.x; if (i
    < d) {
        float val = 0.0f;
        for (int j = 0; j < n; j++) { val +=
            w[i * n + j] * x[j];
        }
        xout[i] = val;
    }
void matmul(float* xout, const float* x, const float* w, int n, int d) {

    // 将数据从主机复制到设备
    cudaMemcpy(d_x, x, n * sizeof(float), cudaMemcpyHostToDevice); cudaMemcpy(d_w,
    w, d * n * sizeof(float), cudaMemcpyHostToDevice);
```

```
    // 配置线程块和线程网格
    int threadsPerBlock = 16;
    int blocksPerGrid = (d + threadsPerBlock - 1) / threadsPerBlock;

    // 启动 CUDA 内核
    matmulKernel<<<blocksPerGrid, threadsPerBlock>>>(d_xout, d_x, d_w, n, d);

    // 将结果从设备复制回主机
    cudaMemcpy(xout, d_xout, d * sizeof(float), cudaMemcpyDeviceToHost);
}
```

执行命令

```
ompute_35 -O3 -std=c++17 -o run run.cu -lm
nvcc warning : The 'compute_35', 'compute_37', 'sm_35', and 'sm_37' architectures are deprecate
d, and may be removed in a future release (Use -Wno-deprecated-gpu-targets to suppress warning)
```

编译通过。

0. 对比分析优化前后的推理性能

```
Once upon a time, there was a little girl named Lily. She loved to sing and dance all day long. One day, she went to her
 to play with it too, so she asked her friend if she could play with it.
Her friend said yes, and they started to play with the toy gun together. They pretended to be cowboys and robbers, shootin
se she was having so much fun.
Suddenly, Lily's little brother came into the room and started to cry. He wanted to play with the toy gun too. Lily rememb
 but you need to say please." Her brother understood and gave the toy gun back to Lily. They both started to play with it
achieved tok/s: 50.750221
```

```
Once upon a time, there was a big train. The train had many cars and a long rail. One day, the train had to separate from
On the rail, there was a delicate flower. The flower was sad because it could not fly. The train stopped and looked at th
The train picked up the delicate flower and put it on a small hill. The flower was happy and started to dance. The train
very day.
achieved tok/s: 20.028612
```

```
Once upon a time, there was a big, powerful elephant named Ella. Ella had a long trunk that she used to spray water in an
One day, Ella met a little bird named Bobby. Bobby was sad because he could not fly. He sat on a tree branch and said, "E
Ella used her trunk to spray water on Bobby. The water made Bobby's wing go up, up, up! Bobby was so happy. He said, "Tha
iends, and they played together every day in the forest.
achieved tok/s: 73.510222
```

```
Once upon a time, there was a little girl named Lily. She loved to play with her toys and color with her crayons. One day,
Lily was so happy and started coloring right away. She enjoyed coloring so much that she didn't want to stop. But her mom
Lily tried to color in different colors, but it wasn't easy. She didn't like getting bored, so she kept coloring. But ther
 color like the other pictures.
Lily got frustrated and threw her crayons on the floor. Her mom told her that it's okay to make mistakes and that she shou
In the end, Lily's coloring book became too messy and ruined. She was sad because she couldn't color anymore and couldn't
achieved tok/s: 33.052495
```

```
Once upon a time, there was a little girl named Lily. She loved playing with her hoop. She would
when she saw a boy who looked very sad.
"What's wrong?" asked Lily.
"I lost my toy car," said the boy. "It was my favorite."
Lily felt sorry for the boy and decided to help him find his toy car. She asked everyone in the p
py and thanked Lily.
Lily felt proud that she could help someone. She went home with a smile on her face, knowing that
achieved tok/s: 7.294999
```


```
Once upon a time, there was a little girl named Lily. She loved playing with her hoop. She would
when she saw a boy who looked very sad.
"What's wrong?" asked Lily.
"I lost my toy car," said the boy. "It was my favorite."
Lily felt sorry for the boy and decided to help him find his toy car. She asked everyone in the p
py and thanked Lily.
Lily felt proud that she could help someone. She went home with a smile on her face, knowing that
achieved tok/s: 7.294999
```


```
Once upon a time, there was a little girl named Lily. She loved to play outside and pick flowers.
p and showed it to her mom.
Her mom looked at the rock and said, "That's a very special rock, Lily. It's called a diamond. It
Later that day, Lily's mom showed her how to print a picture of a flower. Lily thought it was ver
As Lily grew up, she became very popular at school. Everyone wanted to be her friend and play wit
ends so they could all see how special it was.
achieved tok/s: 13.151631
```

如图所示，用CUDA优化后性能提升明显

实验结果

	CPU推理性能	CUDA推理性能
15M	51	72.91
42M	20	32.95
110M	7	12.56

可以看见CUDA优化后`llama2`的推理性能显著提升。

原理分析

从注释中可以看出，`matmul` 函数实现的是一个大小为 d×nd \times n 的矩阵 WW 与维度为 nn 的列向量 xx 的乘法运算。此外，该函数也是模型推理过程中性能瓶颈的主要来源。GPU 的架构设计注重并行计算能力，通常配备数千个 CUDA 核心，这些核心能够同时处理大量线程任务。GPU 在半精度（FP16）、混合精度（FP16+FP32）以及整型（INT8）运算方面具有硬件加速功能。在推理阶段，经常使用低精度计算，这不仅提高了执行速度，还减少了内存带宽的需求。相比于 CPU，GPU 的显存带宽要高得多。例如，现代 GPU 的显存带宽可超过 1 TB/s，而 CPU 的内存带宽通常只有数十 GB/s。更高的显存带宽使 GPU 能够更快速地访问和存储大规模数据，在推理任务中对模型权重及中间计算结果的频繁访问也更加高效。