K230部署 Qwen2.5-0.5B LLM

参考教程:

https://mp.weixin.qq.com/s/HCuQEqD2UzBD65I3elbCow

1. 烧写镜像

- qwen-0.5B例程支持**1GB/2GB**内存01 studio板子
- 01 studio sysimage-sdcard.img

```
$ ls -l sysimage-sdcard.img
-rw-rw-rw- 1 zhangyang zhangyang 3355463680 Mar 20 15:13 sysimage-sdcard.img

$ md5sum sysimage-sdcard.img
30cc3779e40e50fb6a7eda8aa8767962  sysimage-sdcard.img
```

2. 修改/编译程序

> 若用户不需要修改并编译程序, 可跳过本章节, 直接执行章节3

2.1 下载riscv64交叉编译工具链

若不须修改程序, 可跳过此步骤, 直接运行镜像自带的k230 LLM程序

- 下载Xuantie-900-gcc-linux-6.6.0-glibc-x86_64-V2.10.1-20240712.tar.gz
- 解压到/path/to/k230_ai_assistant/src/toolchain目录

```
cd /path/to/k230_ai_assistant/src
wget -c https://occ-oss-prod.oss-cn-
hangzhou.aliyuncs.com/resource//1721095219235/Xuantie-900-gcc-linux-6.6.0-glibc-
x86_64-V2.10.1-20240712.tar.gz
mkdir toolchain
tar xzf Xuantie-900-gcc-linux-6.6.0-glibc-x86_64-V2.10.1-20240712.tar.gz  -C
toolchain/
ls -l toolchain
```

2.2 qwen_chat

```
$ cd k230_ai_assistant/src/qwen_chat/
$ ./riscv64_build.sh
$ ll build/qwen_chat
-rwxrwxr-x 1 zhangyang zhangyang 4339952 Mar 20 14:40 build/qwen_chat*
```

将build/qwen_chat 覆盖到k230 01 studio板子的/app/qwen_chat/qwen_chat即可.

2.3 voice_assistant

```
$ cd k230_ai_assistant/src/voice_assistant
$ ./riscv64_build.sh
$ ls -l build/voice_assistant
-rwxrwxr-x 1 zhangyang zhangyang 9870208 Mar 20 14:43 build/voice_assistant
```

将build/voice_assistant覆盖到k230 01 studio板子的/app/voice_assistant/voice_assistant即可.

3. 上板运行

3.1 qwen_chat

功能: 文本输入, 文本输出

运行方法:

- 执行run.sh启动程序， 由于模型较大, 需要等待一段时间
- 启动成功后, 在"Q:"后输入问题, 支持**中/英文**
- 程序检测到回车后, 开始推理, 终端打印问题答案, 最后输出性能信息等.
- 输入"/exit", 退出程序

> 目前限制最大输入+输出=512个token. 输出过程中, 输入+输出token >512后会自动截断

```
[root@canaan ~ ]#cd /app/qwen_chat/
[root@canaan /app/qwen_chat ]#./run.sh
+ echo 'kpu noc infinite'
kpu noc infinite
+ devmem 0x91103028 32 0x30002
+ devmem 0x91301d0c 32 0
+ devmem 0x91301d8c 32 0
+ devmem 0x91213400 32
0x00000700
+ cat /proc/meminfo
MemTotal:        2026016 kB
MemFree:         1967308 kB
MemAvailable:    1974736 kB
Buffers:            4452 kB
Cached:            17788 kB
SwapCached:            0 kB
Active:            10296 kB
Inactive:          14708 kB
Active(anon):         84 kB
Inactive(anon):     2760 kB
```

```
Active(file):      10212 kB
Inactive(file):    11948 kB
Unevictable:           0 kB
Mlocked:               0 kB
SwapTotal:             0 kB
SwapFree:              0 kB
Dirty:                84 kB
Writeback:             0 kB
AnonPages:          2768 kB
Mapped:             7296 kB
Shmem:                80 kB
KReclaimable:       5700 kB
Slab:              19420 kB
SReclaimable:       5700 kB
SUnreclaim:        13720 kB
KernelStack:        1480 kB
PageTables:          628 kB
SecPageTables:         0 kB
NFS_Unstable:          0 kB
Bounce:                0 kB
WritebackTmp:          0 kB
CommitLimit:     1013008 kB
Committed_AS:      39408 kB
VmallocTotal:   67108864 kB
VmallocUsed:        4964 kB
VmallocChunk:          0 kB
Percpu:              104 kB
CmaTotal:         1888256 kB
CmaFree:          1863032 kB
HugePages_Total:       0
HugePages_Free:        0
HugePages_Rsvd:        0
HugePages_Surp:        0
Hugepagesize:       2048 kB
Hugetlb:               0 kB
+ dst=Qwen2.5-0.5B-Instruct
+ ./qwen_chat Qwen2.5-0.5B-Instruct/config.json
model path is Qwen2.5-0.5B-Instruct/config.json
load tokenizer
tokenizer_type = 3
load tokenizer Done
load Qwen2.5-0.5B-Instruct/llm.kmodel ... Load Module Done!

Q: Hello

A: Hello! How can I assist you today?

################################
 total tokens num  = 25
prompt tokens num  = 15
output tokens num  = 10
  total time = 3.23 s
prefill time = 1.06 s
 decode time = 2.16 s
  total speed = 7.75 tok/s
prefill speed = 14.10 tok/s
```

```
 decode speed = 4.62 tok/s
   chat speed = 3.10 tok/s
###################################
```

Q: 背诵静夜思

A: 当然，我可以为您背诵一首著名的唐代诗人李白的《静夜思》：

床前明月光，
疑是地上霜。
举头望明月，
低头思故乡。

这首诗描绘了夜晚在庭院中看到明亮的月亮时的感受。诗人通过想象月亮照亮地面的情景，并且抬头仰望月亮，然后低下身子去思念自己的家乡。整首诗表达了诗人对远方家乡的深深思念之情。

```
###################################
 total tokens num  = 110
prompt tokens num  = 19
output tokens num  = 91
  total time = 23.51 s
prefill time = 0.96 s
 decode time = 22.54 s
  total speed = 4.68 tok/s
prefill speed = 19.72 tok/s
 decode speed = 4.04 tok/s
   chat speed = 3.87 tok/s
###################################
```

Q: 1+2+3+...+100=?

A: This is an arithmetic series problem, where the first term $a = 1$, common difference $d = 1$, and the last term $l = 100$.

The formula for the sum of an arithmetic series is:

$$ S_n = \frac{n}{2} (a + l) $$

where:
- n is the number of terms,
- a is the first term,
- l is the last term.

Given:
- $a = 1$,
- $l = 100$,

we can calculate n as follows:

$$ n = \frac{l - a}{d} + 1 $$
$$ n = \frac{100 - 1}{1} + 1 $$
$$ n = 99 + 1 $$
$$ n = 100 $$

```
Now we can find the sum using the formula:

\[ S_{100} = \frac{100}{2} (1 + 100) \]
\[ S_{100} = 50 \times 101 \]
\[ S_{100} = 5050 \]

So, the sum of the series from 1 to 100 is 5050.

##################################
 total tokens num  = 303
prompt tokens num  = 26
output tokens num  = 277
  total time = 75.12 s
prefill time = 0.97 s
 decode time = 74.15 s
  total speed = 4.03 tok/s
prefill speed = 26.72 tok/s
 decode speed = 3.74 tok/s
   chat speed = 3.69 tok/s
##################################


Q: /exit
[root@canaan /app/qwen_chat ]#
```