GPU memory release问题研究

1、使用 Numba 库进行CUDA管理

参考：

- Numba官方文档：Numba for CUDA GPUs

  Numba是一个Python的即时编译器，在使用NumPy数组和函数以及循环的代码上效果最好。使用Numba最常见的方式是通过它的装饰器集合，这些装饰器可以应用于你的函数，以指示Numba对其进行编译。当调用一个Numba装饰的函数时，它被编译成机器代码 "及时 "执行，随后你的全部或部分代码可以以本地机器代码的速度运行.

keras 的资源释放方式：

```
from keras import backend as K
K.clear_session()
```

Numba的资源管理方式：

```
# 1
from numba import cuda
cuda.select_device(0)
cuda.close()
# User can then create a new context with another device.
cuda.select_device(1)  # assuming we have 2 GPUs
```

```
# 2
from numba import cuda
device = cuda.get_current_device()
device.reset()
```