(3d\_process) brucesz@LAPTOP-26R7SA9G:/data/git/cuda/chapter5/assignment5$ python test\_bert.py

<class 'list'>

input nmae is: input\_ids

onnx-runtime cost: 0.02045726776123047

normal model infer cost: 0.05720186233520508

Saving inputs and output to case\_data.npz ...

tensor([[0, 1, 2, 3, 4, 5, 6, 7, 8]], dtype=torch.int32)