TestBroadcast_TwoAxis

March 19, 2024

[ ]:
```
import torch
torch.cuda.is_available()
!export LC_ALL="en_US.UTF-8"
!export LD_LIBRARY_PATH="/usr/lib64-nvidia"
!export LIBRARY_PATH="/usr/local/cuda/lib64/stubs"
!ldconfig /usr/lib64-nvidia
```

```
/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a symbolic
link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic link
```

[ ]:
```
import torch
def fn(x, y, bsz, num_head, seq_len_1, seq_len_2, head_dim):
    return torch.reshape(x, (bsz, num_head, seq_len_1, seq_len_2, head_dim)) +␣
 ↪torch.reshape(y, (1, num_head, seq_len_1, 1, head_dim))
fnc = torch.compile(fn)
bsz = 4
num_head = 32
seq_len_1 = 64
seq_len_2 = 32
head_dim = 128
seq_len = int(seq_len_1 * seq_len_2)
x = torch.randn([bsz, num_head, seq_len, head_dim]).cuda()
y = torch.randn([1, seq_len, head_dim]).cuda()
z = fnc(x, y, bsz, num_head, seq_len_1, seq_len_2, head_dim)
print(z[0,0,0,0,0])
```

```
tensor(2.1907, device='cuda:0')
```

```
[ ]: !cat "/tmp/torchinductor_root/yc/
      ↪cyctp2zfav4uedhth2nudcyrf3tqqratlmhgixbpctnrhucibrut.py"
```

```
from ctypes import c_void_p, c_long
import torch
import math
import random
import os
import tempfile
from math import inf, nan
from torch._inductor.hooks import run_intermediate_hooks
from torch._inductor.utils import maybe_profile
from torch._inductor.codegen.memory_planning import _align as align

from torch import device, empty, empty_strided
from torch._inductor.codecache import AsyncCompile
from torch._inductor.select_algorithm import extern_kernels

aten = torch.ops.aten
inductor_ops = torch.ops.inductor
assert_size_stride = torch._C._dynamo.guards.assert_size_stride
alloc_from_pool = torch.ops.inductor._alloc_from_pool
reinterpret_tensor = torch.ops.inductor._reinterpret_tensor
async_compile = AsyncCompile()


# kernel path: /tmp/torchinductor_root/s2/cs2cs5gcqipxcz74dyzrqnudrdggcuuhjblfwo
rul5dgofzq5mi6.py
# Source Nodes: [add], Original ATen: [aten.add]
# add => add
triton_poi_fused_add_0 = async_compile.triton('triton_', '''
import triton
import triton.language as tl
from torch._inductor.ir import ReductionHint
from torch._inductor.ir import TileHint
from torch._inductor.triton_heuristics import AutotuneHint, pointwise
from torch._inductor.utils import instance_descriptor
from torch._inductor import triton_helpers

@pointwise(
    size_hints=[33554432],
    filename=__file__,
    triton_meta={'signature': {0: '*fp32', 1: '*fp32', 2: '*fp32', 3: 'i32'},
'device': 0, 'device_type': 'cuda', 'constants': {}, 'configs':
[instance_descriptor(divisible_by_16=(0, 1, 2, 3), equal_to_1=(),
ids_of_folded_args=(), divisible_by_8=(3,))]},
```

```
    inductor_meta={'autotune_hints': set(), 'kernel_name':
'triton_poi_fused_add_0', 'mutated_arg_names': []},
    min_elem_per_thread=0
)
@triton.jit
def triton_(in_ptr0, in_ptr1, out_ptr0, xnumel, XBLOCK : tl.constexpr):
    xnumel = 33554432
    xoffset = tl.program_id(0) * XBLOCK
    xindex = xoffset + tl.arange(0, XBLOCK)[:]
    xmask = xindex < xnumel
    x4 = xindex
    x0 = xindex % 128
    x2 = (xindex // 4096) % 2048
    tmp0 = tl.load(in_ptr0 + (x4), None)
    tmp1 = tl.load(in_ptr1 + (x0 + (128*x2)), None,
eviction_policy='evict_last')
    tmp2 = tmp0 + tmp1
    tl.store(out_ptr0 + (x4), tmp2, None)
''')

import triton
import triton.language as tl
from torch._inductor.triton_heuristics import grid, start_graph, end_graph
from torch._C import _cuda_getCurrentRawStream as get_cuda_stream


async_compile.wait(globals())
del async_compile

def call(args):
    arg0_1, arg1_1 = args
    args.clear()
    assert_size_stride(arg0_1, (4, 32, 2048, 128), (8388608, 262144, 128, 1))
    assert_size_stride(arg1_1, (1, 2048, 128), (262144, 128, 1))
    with torch.cuda._DeviceGuard(0):
        torch.cuda.set_device(0) # no-op to ensure context
        buf0 = empty((4, 32, 64, 32, 128), device='cuda', dtype=torch.float32)
        # Source Nodes: [add], Original ATen: [aten.add]
        stream0 = get_cuda_stream(0)
        triton_poi_fused_add_0.run(arg0_1, arg1_1, buf0, 33554432,
grid=grid(33554432), stream=stream0)
        del arg0_1
        del arg1_1
        return (buf0, )


def benchmark_compiled_module(times=10, repeat=10):
    from torch._dynamo.testing import rand_strided
```

```
    from torch._inductor.utils import print_performance
    arg0_1 = rand_strided((4, 32, 2048, 128), (8388608, 262144, 128, 1),
device='cuda:0', dtype=torch.float32)
    arg1_1 = rand_strided((1, 2048, 128), (262144, 128, 1), device='cuda:0',
dtype=torch.float32)
    fn = lambda: call([arg0_1, arg1_1])
    return print_performance(fn, times=times, repeat=repeat)


if __name__ == "__main__":
    from torch._inductor.wrapper_benchmark import compiled_module_main
    compiled_module_main('None', benchmark_compiled_module)
```