note of 《FAST CONVOLUTIONAL NETS WITH fbfft : A GPU PERFORMANCE EVALUATION》

zxp

November 2, 2024

1 content

FAST CONVOLUTIONAL NETS WITH fbfft: A GPU PERFORMANCE EVALUATION is a 15-year-old paper written by Facebook that implements the open source Facebook fbcdnn and part of fbcdnn for FFT-based convolution and analyzes the cufft library.

The paper also points out that the reason for expanding matrices into convolution is that matrix multiplication is a basic operation of linear algebra that is well optimized on almost any platform.

The paper mentioned the data structure, using the regular BDHW, which was converted to WHDB after the FFT conversion to use the cgeem library (not knowing the details of the FFT, not knowing why). for larger batches over small matrices, the cublasCgemmBatched library call; For larger batches over small matrices, the Cublascgemmbatched library call; • for smaller batches over larger matrices, multiple cublasCgemm calls from the host; • for intermediate batch and matrix sizes, devices of compute capability 3.5 and higher sup-port dynamic parallelism which allows CUDA kernels to launch other kernels. This can bebeneficial for many launches over small matrices. According to the paper, cuFFT implements FFTs with the ubiquitous Cooley-Tukey algorithm (Cooley & Tukey (1965))which takes advantage of trigonometric equalities to recursively decompose and reuse computations.

The FFT library implemented in this paper is filled implicitly when loading data, while cufft is explicit, so the memory required by fbfft is smaller than that of cufft library.

2 Feelings

This paper open source FFT implementation, maybe can refer to