CULLM THEORY

ERNEST YEUNG FOR IN SERVICE OF X LLC INSERVICEOFXY@PROTON.ME

Contents

Abstract. Your abstract text here.

1. Introduction

2. LLM

2.1. Attention.

2.1.1. *Attention Forward.* Consider in `https://github.com/karpathy/llm.c/blob/7ecd8906afe6ed7a2b2cdb731c042f26d525b820/dev/cuda/attention_forward.cu#L160` `attention_query_key_kernel1` what this means mathematically:

For $k_x = i_x + j_x M_x = 0 \dots N_x M_x - 1$, where $i_x = 0 \dots M_x - 1$ and $j_x = 0 \dots N_x$, corresponding to total number of threads, thread index, and block index, respectively, if for

$$\begin{aligned} B &\equiv \text{ Batch size} \\ NH &\equiv \text{ Number of attention heads} \\ T &\equiv \text{ Sequence length} \end{aligned}$$

then if we required total number of threads $= B * NH * T * T$, then

Consider `https://github.com/karpathy/llm.c/blob/7ecd8906afe6ed7a2b2cdb731c042f26d525b820/dev/cuda/attention_forward.cu#L192` `attention_softmax_kernel1`

Date: November 27, 2024.
Key words and phrases. LLM, Attention, CUDA.