https://www.douyin.com/video/7344678214398151946

# 标题:未找到标题  
## 关键字: 未找到关键字  
## 作者: 严伯钧  
## 视频ASR文本:  
 这两天呢我在阿开看到了一个清华大学的重要科研成果啊就是清华大学摇班的科学家呢做出了一个关于矩阵乘法的重大突破这个文章足足八十七页长啊这里面作者居然还有个本科生哎这年头的本科生啊真是一届比一届厉害了啊 我就纳闷了这么重要一个计算机算法领域的成果怎么就没有激起什么水花呢没事啊你们不关注我来替你们关注关注老规矩啊塞到 t x y z 里面甭管多长的文章一样盘你问几次就给你盘清楚了啊这篇文章虽然长但很多呢是证明推导过程主要核心啊 其实就是用计算机啊去做矩阵乘法的效率呢又得到了一个提升那咋回事呢这里面逻辑有点长啊你听我慢慢给你讲当然没有耐心听我说的你们自己去 t x y z 找这篇文章叫做 faster matrix multiplication via asymmetric hash 啊就是通过非对称哈西实现的更快 的矩阵乘法啊自己问 ai 去啊那么所以啥是矩阵乘法呢首先啥是矩阵啊矩阵呢是啥我就不多讲了啊这一讲就太长了直接要把大学里面现行代数讲一遍我就假设大家知道什么是矩阵啊比方一个二维方阵 a 它有四个元素从左到右从上到下 分别是 a 一 a 一二 a 二一 a 二二啊乘以另外一个二为方份 b 四个元素分别是 b 一 b 一二 b 二一 b 二二问 a 乘以 b 乘出来的这个新矩阵 c 里面也是四个元素这四个元素分别等于啥啊 那么这个矩阵乘法的定义很简单分别是 c 一等于 a 一乘以 b 一加 a 二乘以 b 二二一 c 二等于 a 一乘以 b 一加 a 二乘以 b 二一 c 二等于 a 二一乘以 b 一二加 a 二乘以 b 二二那推广一下对于一个 n 为矩阵 a 乘以另外一个 n 为矩阵 b 乘出来的这个新矩阵 c 啊第二 i 行 d j 列的元素 c i j 就等于矩阵 a 的 d i 行和矩阵 b 的 d j 列把这个 a 矩阵的 i 行和 b 矩阵的 j 列当成两个矢量做点乘那就是 c i j 等于求和 a i k 乘以 b k j k 从 k 等于一加到 k 等于 n 啊 那这个呢是矩阵乘法的定义你别看这个复杂啊但里面呢就是做很多次乘法然后再做很多次的加法那我们算一下一个 n 乘以 n 的矩阵乘法里面要算多少次乘法和多少次加法 简单算一下呢就知道这里面乘法要做 n 的三次方次而加法要做 n 的平方乘以 n 减一次啊 那这个数字有啥意义呢哎对计算机来说那就有意义了因为对于计算机来说加法是很简单的但是乘法却没有那么简单所以一个矩阵乘法的运算如果乘法次数少速度就会快而对于现在的 ai 来讲啊主要是神经网 网络的 ai 里面做的计算就主要是矩阵乘法因为每个神经元里面的这个 w 参数它就是个矩阵数据呢是以矢量的方式进入神经元然后要算的就是个矩阵乘法如果这个乘法算的速度特别快 那么 ai 运算的效率就会高那有人就要问了你都知道矩阵乘法的公式了直接带入算不就完了吗没那么简单啊因为当矩阵的维度很大也就是 n 很大的时候你会发现这个差别它就大了去了 比方现在的大语言模型神经元里的矩阵都挺大的比方说 n 等于五百的话那么做乘法的次数如果是 n 的三次方那就是要做一点二五亿次的乘法如果我们能让矩阵乘法当中计算机做乘法的次数变成 n 的平方那你效率就直接提升五百倍 好了你可能要问这个矩阵乘法的定义都清清楚楚的写在那里的那就是 n 的三次方呀咋还能算乘法的次数比这个少呢哎还真可以这就是算法的领域了 最早是一九六九年的时候有个学者叫 stressen 啊他提出个 stressen 算法比方一个二乘二的矩阵正常如果去按照定义做乘法我要做的是二的三次方也就是八次乘法但是 stressen 这个算法呢就只要做七次就可以了怎么做的呢他就是这么搞的就先搞七个量啊我就写在这我就不念了太长了啊 这七个量每个量里都只有一次乘法对吧好了那乘出来的新矩阵里那就没有乘法了新矩阵的这四个元素就是分别 c 一等于 p 一加 p 四减 p 五加 p 七 c 二等于 p 三加 p 五 c 二等于 p 二加 p 四 c 二等于 p 一加 p 二减 p 三加 p 六 这里只有加法是不是就少做了一次乘法呀这还是二乘二的矩阵那把这个办法推广到 n 乘以 n 的矩阵按照这个 stress 的办法可以把这个乘法次数变成 n 的二点八一次方那如果是个五百乘五百的矩阵那差不多就五百的这个零点一九次方那效率就是原来的三倍所以你看到了啊用新的算 法是可以把这个乘法次数逐渐减少的那最少多少次呢哎 n 的平方次因为你的新矩阵要得出 n 的平方个元素所以至少得做 n 的平方次乘法吧乘法就不能比这个更少了当然特殊的矩阵除外啊比方你先分析下这个矩阵有没有什么特殊性质比方有很多元素是零那就可以乘的次数那就更少 但一般来说 n 的平方次乘法是理论最小值然后呢这个领域就一直在发展到一九八一年的时候有一个叫 shanghai 的学者弄出来了个 laser 算法这个激光法啊 总的意思呢就是把这个矩阵乘法搞到了什么张亮分析里面变成了另外一个问题他可以把这个矩阵乘法的次数变成 n 的二点五二二次方啊这样一下呢如果是个五百位的矩阵那就效率他就又高了五点六倍 然后呢沿着这个激光法呀之前的世界纪录是二点三七二八六零哎那终于要说到清华大学这篇文章了清华这篇文章啊是用了一种新的办法就是这个 反哈西的办法结合了机器学习把这个世界纪录呢又给刷新了刷到多少呢二点三七一八六六啊差不多减少了零点零零一如果是个五百为的矩阵呢效率大概提升了百分之一但不要小看这个零点零零一的提升啊这次主要的创新呢是一种方法论的创新 哈佛有个著名学者评论啊这是这十年来在这个方向上最大的进展当然具体是什么我自己也没有完全搞明白因为这里面的数学和计算机知识实在太深了人家搞了几十年的东西我一下也不是搞得很清楚当然通过问题 x y z 还是可以逐渐学习清楚的 不用上传论文直接在这个框里问问题他不光给你解释还可以给你找相关的资料并且把相关资料总结好以后告诉你为什么要看这些资料哎并且还能建议你问什么问题就这么个学法相信再复杂的概念很快也就搞明白了搞学术的奔走相告啊啥也别说了为清华点个赞