背景

- 现在有可信的CPU和不可信的GPU，怎么高效的完成大模型的各种操作?

目录

- 需要进行的操作
- 威胁模型
- 如何加密

矩阵操作

- 非线性操作一定在CPU完成。
- 对于线性操作，分为以下两种。
  - 密文乘明文，例如各种投影操作
  - 密文乘密文，例如 $Q \cdot K^T$，$S \cdot V$，反向传播应该也有。

- 目标：把计算密集的部分放在GPU完成，访存瓶颈可以由CPU算。

威胁模型

- 不安全的GPU可能怎么进行攻击?
- 攻击者类型为**半诚实**的:
  只能在通信信道上截获密文序列，无法干预或修改通信流程。
  输入数据分布、加密算法结构及参数对攻击者完全公开。
- **攻击者目标**是尽可能获取特征向量的信息。

- 攻击者能力有
  - 可截获任意数量的密文，但无法得到对应的明文 。
  - **不存在已知明文或选择明文查询**；攻击者亦不可将任意明文提交给加密器获得密文。
  - 攻击者**不能通过重放已截获的密文**来诱导目标环境产生额外反馈或信息。
  - 可执行任意线性代数运算和统计分析，但不持有任何秘密。

思路

- 对于密文乘明文 $c(X)Y$，考虑加性加密，注入加性噪声对抗统计攻击。
- 即 $Z = XY = (X + \mathbf{n})Y - \mathbf{n}Y$，$\mathbf{n}$为随机噪声向量。

- 对于密文乘密文$c(X)c(Y)$，考虑加入一个旋转矩阵 R，$R^{-1} = R^T$。
- $Z = XY = XY^{TT} = XRR^{-1}Y^{TT} = (XR)(Y^T R^T)^T$

- actfn(XW)

挑战

- $\mathbf{n}Y$ 在CPU上的计算开销
- XR 在CPU上的计算开销

密文乘明文

简单想法

- 注意到明文Y都是权重，是不变的
- 因此可以准备t个噪声向量 $\mathbf{n}_0, \mathbf{n}_1, \ldots, \mathbf{n}_{t-1}$, $t > r$ (r为激活值向量的秩)
- 预计算出 $\Delta z_i = \mathbf{n}_i Y$。每次随机选取一个 i，加密就是$c_x = x + \mathbf{n}_i$ ，解密就是 $z = c(z) - \Delta z_i$。
- 为了掩盖x，选取一个缩放系数让 x, $\mathbf{n}_i$ 能量（长度）相当即可。运行时复杂度 $O(d)$

- 问题：对$c(x)$聚类大概率能在超球面上看到 r 个凸起
  泄漏了噪声方向信息，进而泄漏 x 的信息

噪声向量旋转

- 每次选取随机i, j
- 随机的系数α β
- $c(x) = x + \alpha \mathbf{n}_i + \beta \mathbf{n}_j$，$z = c(z) - \alpha \Delta z_i - \beta \Delta z_j$。这样就有$\frac{t(t-1)}{2}$种组合，只要不是特别小就可以抵抗攻击。

- 更强的加密：考虑每次使用噪声后，把两个向量随机混合旋转出两个新向量作为 $\mathbf{n}_i'$ $\mathbf{n}_j'$，替代原有噪声向量。
- 和上面一样，新的Δz_i Δz_j可以直接求出，不需要再做一次矩阵乘法。
- 这样两条不同的密文的噪声部分从来不会共享同一个扰动向量或比值固定的方向成分，从而抵抗攻击。
- 甚至可以乘以系数normalize $c(x)$，在统计意义上是非线性的，彻底破坏了长度信息，扰动了角度信息。

一些细节

- 噪声向量的选取
  - 选取若干明文数据进行推理，记录其激活值，进行部分奇异值分解，记录其显著的基向量和对应方向的能量。基向量之间进行若干次上述混合操作即为噪声向量。
    明文：拆成两部分给GPU算，GPU难以组合。代价是算力翻倍。
  - 这样可以保证激活值每个显著方向的信噪比都低。

- 保持分量的旋转混合操作
  - 为了保持噪声向量不同分量的长度，生成的新的噪声向量的系数要满足下面的条件

$$n_1' = an_1 + bn_2$$
$$n_2' = cn_1 + dn_2$$

  - $n_1^2 = n_1' \cdot n_1 + n_2' \cdot n_1$
    $n_2^2 = n_1' \cdot n_2 + n_2' \cdot n_2$
  - 随便搞搞就解出来了

$$A = n_1 \cdot n_1$$
$$B = n_1 \cdot n_2$$
$$C = n_2 \cdot n_2$$
$$a, b \in \mathbb{R}*$$
$$c = \frac{C(A-B)}{AC-B^2} - a$$
$$d = \frac{A(C-B)}{AC-B^2} - b$$

密文乘密文

- 考虑下面两种特殊的旋转矩阵
- 矩阵由对角线上的若干小的 $n \times n$ 旋转矩阵构成，$n \ll d$
- $$R_s = \begin{pmatrix} R' & & \\ & R' & \\ & & \ddots \\ & & & R' \end{pmatrix}$$
- xR_s等价于计算若干个$x'R'$，时间复杂度从 $O(d^3)$ 变为了$O(dn^2)$

- 矩阵是一个置换矩阵，具体来说，每一行每一列只有一个元素是1，其余全为0。
- 其等价于把输入元素按照某种全排列 $\sigma(i)$ 重新排列。
- $R_{p_\sigma} = \left(p_{ij}\right)_{1 \le i,j \le n}, \quad p_{ij} = \delta_{j,\sigma(i)}$
- xR_{p_σ}等价于把 x_i按照 $\sigma(i)$ 重新排列，复杂度 $O(d)$

因此，只需要对 x_1, x_2明文应用若干次以上两种旋转矩阵，就可以得到点积不变的密文。

如果要加密点积，可以考虑再乘上随机标量或者旋转矩阵，即 $c(x) = \lambda R x R_1 R_2 \ldots R_n$