Full Presentation Title

Optional Subtitle

authors

November 17, 2025

Outline

1. Introduction

2. More Content

Slide Title

This is your first content slide.

- Point 1
- Point 2
- Point 3

Simplified Foundation Model

$$y = f_n(f_{n-1}(...f_1(x)))$$

$$f_i = \sigma(W_i x + b_i)$$

Our input x is a 2d matrix, the W's are linear transformations so they can be represented as matrices, and σ is a nonlinear function.

Tensor Parallelism

Tensor Parallelism(Column)

We can split the weight matrix W into smaller matrices depending on the number of devices we have. Each W_i^j is stored on a different device, and along with it we send the entries of x that correspond to the columns of W_i that are held within each W_i^j as a column vector, x_{L_j}.

$$W_i = \begin{bmatrix} W_i^1 & W_i^2 & \dots & W_i^m \end{bmatrix}$$

Each device is then computing $W_i^j x_{L_j}$ which we can think of as

$$W_i^j x_{L_j} = \begin{bmatrix} w_{11}^j & w_{12}^j & \dots & w_{1k}^j \\ w_{21}^j & w_{22}^j & \dots & w_{2k}^j \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1}^j & w_{m2}^j & \dots & w_{mk}^j \end{bmatrix} \begin{bmatrix} x_p \\ x_p + 1 \\ \vdots \\ x_p + k - 1 \end{bmatrix}$$

Tensor Parallelism(Column) pt 2

Thank You

Questions?