Full Presentation Title

Optional Subtitle

authors

November 17, 2025

Outline

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.

Pipeline Parallelism

Tensor Parallelism(Row)

We can split the weight matrix W_i intro groups of rows depending on how many devices we have. Each entry of the kth result of $W_i x$ is equal to the kth row of W_i times x, so we just have each device compute $C_j x$. Since each device is fully responsible for specific entries of the result, we can also give the corresponding entries of b_i, which we will label b_{L_j} to each device and apply our nonlinear function. We end up with each device computing $\sigma(C_j x + b_{L_j})$, and then gathering the various elements from each device into a single vector to be fed into the next function

$$W_i = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_m \end{bmatrix}$$

Tensor Parallelism(Column)

We can split the weight matrix W_i into smaller matrices depending on the number of devices we have. Each A_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 A_j as a column vector, x_{L_j}.

$$W_i = \begin{bmatrix} A_1 & A_2 & \dots & A_3 \end{bmatrix}$$

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

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

Tensor Parallelism(Column) pt 2

Thank You

Questions?