

Master note

Components needed

1. Communication layer
2. Linear algebra operation
3. Autodiff system
4. Optimizer
5. Neural network block
6. TUI

Neural network block

Attention

$$Q = W_q X + b_q \mathbf{1}^T$$

$$K = W_k X + b_k \mathbf{1}^T$$

$$V = W_v X + b_v \mathbf{1}^T$$

$$S = K^T Q$$

$$\tilde{V} = V \cdot \text{softmax}\left(\frac{S}{\sqrt{d_{\text{attn}}}}\right)$$