

Transformer Optimization Research

Flash Attention: $O(n)$ memory complexity