Multi-Head Attention: Parallel Computation

Input (512 × 768)

Head 1
Q/K/V (64 dims)
Attention 512×512
Output (512×64)
64 dimensions per head

Head 2
Q/K/V (64 dims)
Attention 512×512
Output (512×64)

Head 3
Q/K/V (64 dims)
Attention 512×512
Output (512×64)

...

Head 12
Q/K/V (64 dims)
Attention 512×512
Output (512×64)

Parallel execution on GPU

12 heads × 64 dims = 768 total

Concatenate
12 × 64 = 768 dims

Output (512 × 768)

Same total computation as single head