

BERT-Base: Parameter and Memory Breakdown

Parameter Distribution (110M Total = 440 MB)

Feed-Forward Networks: 57M parameters (52%)

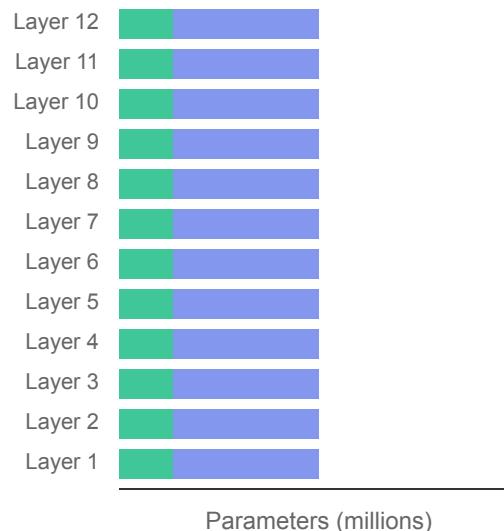
Embeddings: 23M (21%)

Attention: 21M (19%)

Other: 9M

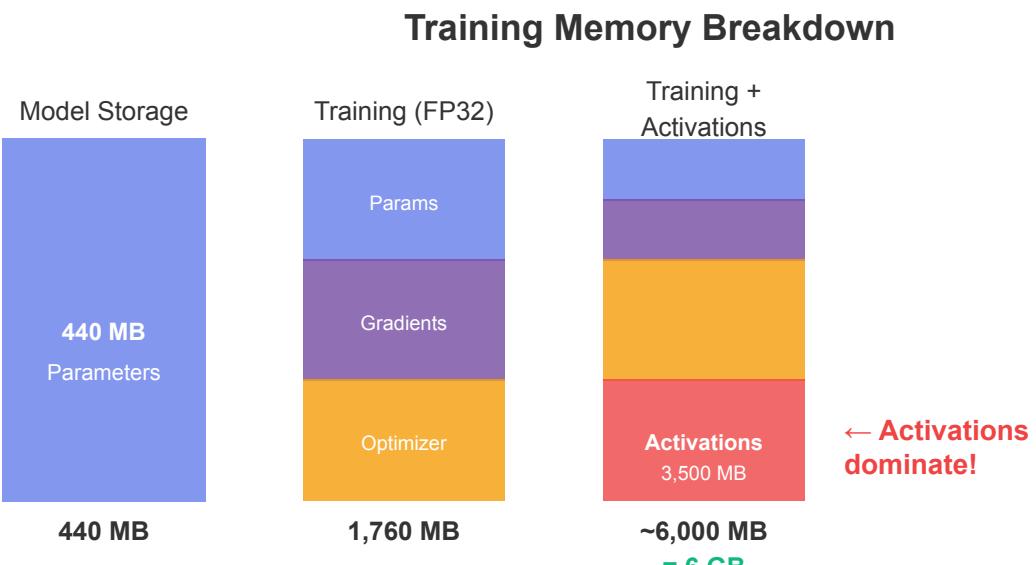
0M 55M 110M

Per-Layer Breakdown (12 Layers)



Attention: 1.77M
Feed-forward: 4.72M
Each layer: ~6.5M parameters

Parameters (millions)



← Activations dominate!

Key Insights for Leaders

- Feed-forward layers: 52% of parameters
- Activations: 60% of training memory
- Doubling batch size ≈ doubles memory
- Doubling sequence length ≈ doubles memory

- Optimizer states: 2× parameter memory
- Total training memory >> model size
- Memory is often the limiting factor
- Gradient checkpointing trades speed for memory