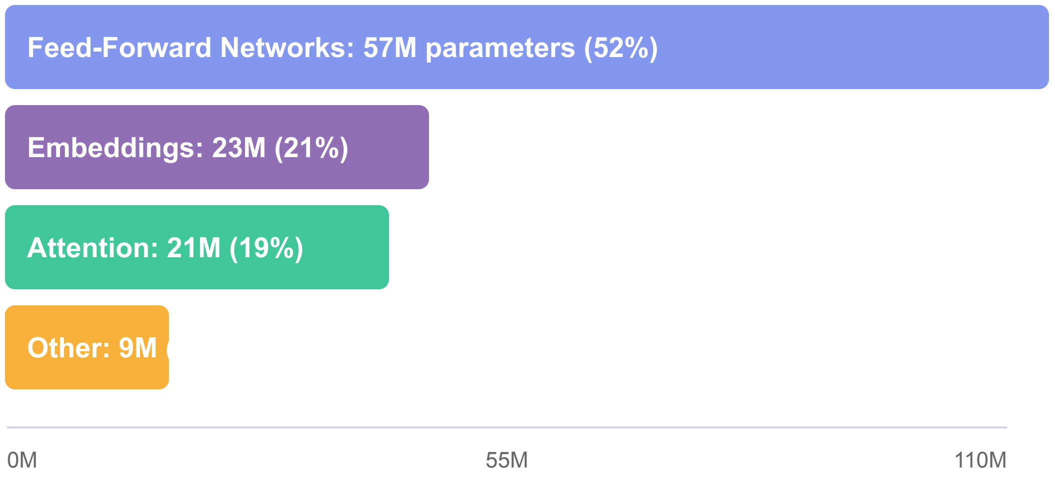
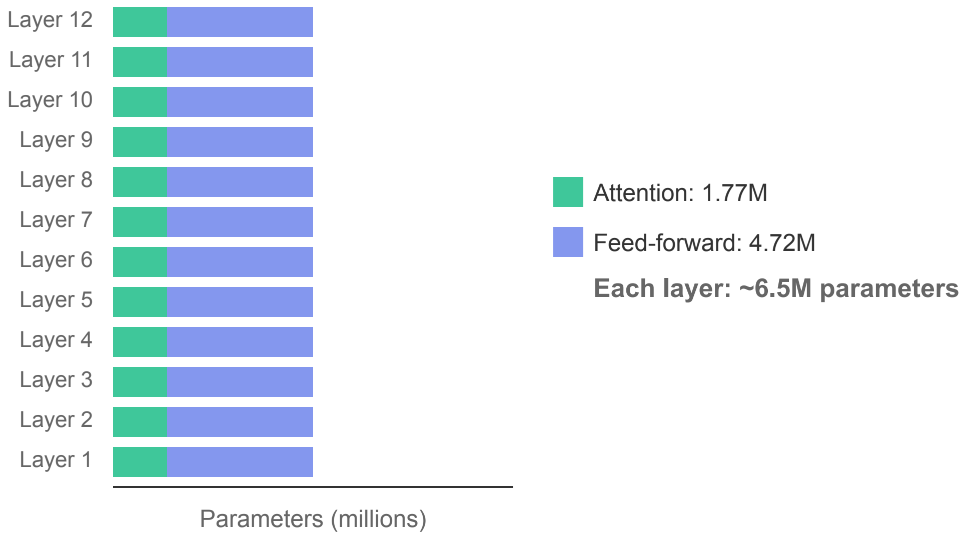
BERT-Base: Parameter and Memory Breakdown

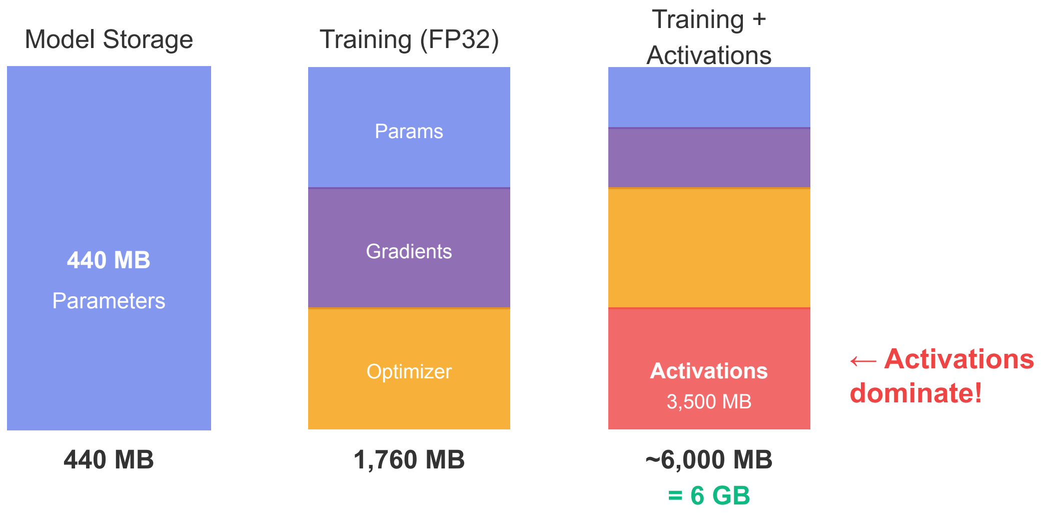
Parameter Distribution (110M Total = 440 MB)



Per-Layer Breakdown (12 Layers)



Training Memory Breakdown



Batch size = 32, Sequence length = 512

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