# FP32 Training Memory Analysis for ResNet-50 and DeiT-Base / ViT-B/16

Subhasish Bandyopadhyay

April 2024

## 1 Introduction

This document analyzes the FP32 training memory requirements for ResNet-50 and DeiT-Base / ViT-B/16 models at 224×224 resolution with a batch size of 64, using the AdamW optimizer. The analysis focuses on saved activations (tensors retained for backward at the end of forward) and parameter-side memory (weights, gradients, first/second moments). Transient workspaces are excluded and reported as a separate headroom band. Units are defined as FP32 = 4 bytes/element, 1 MiB = $2^{20}$ bytes, and 1 GiB = $2^{30}$ bytes.

## 2 ResNet-50: Saved-Activation Geometry and Totals

For a bottleneck block at spatial size $H \times W$ with inner channels $P$ and expansion factor 4, the main-path outputs retained per block are:

$$\underbrace{(P \cdot H \cdot W)}_{1\times1 \text{ reduce}} + \underbrace{(P \cdot H \cdot W)}_{3\times3} + \underbrace{(4P \cdot H \cdot W)}_{1\times1 \text{ expand}} = 6\,P\,H\,W$$

For the first block of each stage, the skip path includes a 1×1 projection, retaining the previous stage's output for weight gradient computation. Batch size $B = 64$. Element counts are per image, multiplied by $B$, then by 4 bytes/element.

### 2.1 Stage-wise Accounting

- **conv1 out** ($64 \times 112 \times 112$): $64 \times 112 \times 112 \times 64 \times 4/2^{20} = 196.00$ MiB
- **maxpool out** ($64 \times 56 \times 56$): $64 \times 56 \times 56 \times 64 \times 4/2^{20} = 49.00$ MiB
- **Stage 1** ($56 \times 56$, $P = 64$, 3 blocks): Main path = $3 \times 6 \times 64 \times 56 \times 56 \times 64 \times 4/2^{20} = 882.00$ MiB; downsample input = $64 \times 56 \times 56 \times 64 \times 4/2^{20} = 49.00$ MiB $\rightarrow$ 931.00 MiB
- **Stage 2** ($28 \times 28$, $P = 128$, 4 blocks): Main path = $4 \times 6 \times 128 \times 28 \times 28 \times 64 \times 4/2^{20} = 384.00$ MiB; downsample input = $256 \times 56 \times 56 \times 64 \times 4/2^{20} = 400.00$ MiB $\rightarrow$ 784.00 MiB
- **Stage 3** ($14 \times 14$, $P = 256$, 6 blocks): Main path = $6 \times 6 \times 256 \times 14 \times 14 \times 64 \times 4/2^{20} = 301.00$ MiB; downsample input = $512 \times 28 \times 28 \times 64 \times 4/2^{20} = 238.00$ MiB $\rightarrow$ 539.00$MiB$**Stage 4**($7 \times 7$, $P = 512$, 3 blocks): Main path = $3 \times 6 \times 512 \times 7 \times 7 \times 64 \times 4/2^{20} = 88.50$ MiB; downsample input = $1024 \times 14 \times 14 \times 64 \times 4/2^{20} = 70.75$ MiB $\rightarrow$ 159.25$MiB$

Total saved activations (excluding indices):

$$196.00 + 49.00 + 931.00 + 784.00 + 539.00 + 159.25 = \mathbf{2658.25}\ \text{MiB} = \mathbf{2.596}\ \text{GiB}$$

Max-pool indices ($64 \times 56 \times 56$, 32-bit): $64 \times 56 \times 56 \times 64 \times 4/2^{20} = 49.00$ MiB.

### 2.2 Parameter-side Memory

Parameters $\approx 25.6$ M, 16 bytes/parameter (weights, gradients, first/second moments): $25.6 \times 10^6 \times 16/2^{30} = 0.381$ GiB.

### 2.3 ResNet-50 FP32 Training Memory (Excluding Workspace)

- Saved activations: 2.596 GiB
- Pool indices: 0.048 GiB
- Parameters + gradients + moments: 0.381 GiB
- **Total**: $\mathbf{3.03}$ GiB

Practical headroom (transient workspaces, allocator): +1020% $\rightarrow$ 3.333.63 GiB.

## 3 DeiT-Base / ViT-B/16: Saved-Activation Geometry and Totals

Patch size 16 on $224 \times 224$ yields $14 \times 14 = 196$ patch tokens. With a CLS token, sequence length $S = 197$. Hidden size $H = 768$, heads $h = 12$, layers $L = 12$, batch $B = 64$. Define $BSH = B \cdot S \cdot H = 9{,}682{,}944$.

Retained tensors per transformer layer (upper-bound for vanilla module graphs):

- Residual checkpoints and Q/K/V projections + MLP hidden: $9\,BSH$
- Attention probabilities $(B, h, S, S)$ only for naïve MHA that materializes softmax probabilities. Modern SDPA/FlashAttention recomputes in backward, not retaining $B\,h\,S^2$.

A single $BSH$ patch-embedding output is retained once (for initial projection/pos-embed path).

### 3.1 Baseline A Naïve MHA (Stores $B\,h\,S^2$)

Per layer: $9\,BSH + B\,h\,S^2$.

- Per-layer saved activations: $(9 \times 9{,}682{,}944 + 64 \times 12 \times 197^2) \times 4/2^{20} = 446.14$ MiB
- Patch-embed saved once: $9{,}682{,}944 \times 4/2^{20} = 36.94$ MiB
- 12-layer saved activations: $(12 \times 446.14 + 36.94)/1024 = 5.264$ GiB

### 3.2 Baseline B SDPA / FlashAttention (No $B\,h\,S^2$)

Per layer: $9\,BSH$.

- Per-layer saved activations: $9 \times 9{,}682{,}944 \times 4/2^{20} = 332.44$ MiB
- Patch-embed saved once: 36.94 MiB
- 12-layer saved activations: $(12 \times 332.44 + 36.94)/1024 = 3.932$ GiB

### 3.3 Parameter-side Memory

Parameters $\approx 86.57$ M, 16 bytes/parameter: $86.57 \times 10^6 \times 16/2^{30} = 1.290$ GiB.

### 3.4 ViT-B/16 FP32 Training Memory (Excluding Workspace)

- **Naïve MHA**: Saved activations 5.264 GiB + params 1.290 GiB $\rightarrow$ $\mathbf{6.55}$ GiB
- **SDPA/Flash**: Saved activations 3.932 GiB + params 1.290 GiB $\rightarrow$ $\mathbf{5.22}$ GiB

Practical headroom: +1020%:

- Naïve MHA: 7.217.86 GiB
- SDPA/Flash: 5.746.27 GiB

Minor additions: LayerNorm statistics are $O(BS)$ (kB-scale); dropout masks (if uint8) add $\approx 910$ MiB per $BSH$ tensor using dropout. Both are sub-5% at $S = 197$.

# 4 Side-by-Side Totals (FP32, AdamW, Excluding Workspace)

| Model | Saved Activations | Indices / Masks | Params + Grads + Moments | Total |
|---|---|---|---|---|
| ResNet-50 | 2.596 GiB | 0.048 GiB | 0.381 GiB | 3.03 GiB |
| ViT-B/16 Naïve MHA | 5.264 GiB | ∼small | 1.290 GiB | 6.55 GiB |
| ViT-B/16 SDPA/Flash | 3.932 GiB | ∼small | 1.290 GiB | 5.22 GiB |

Table 1: FP32 training memory comparison

**Ratios (ViT : ResNet, totals):**

- Naïve MHA: $6.55/3.03 \approx 2.17\times$
- SDPA/Flash: $5.22/3.03 \approx 1.73\times$

# 5 Implementation Notes

1. Gradient checkpointing reduces saved-activation residency by $\approx$ 3050% with added compute; applies to both CNN and transformer blocks.
2. Attention kernel choice governs whether $B\,h\,S^2$ appears in the live set. SDPA/FlashAttention removes this term; naïve MHA retains it.
3. Operator fusion/compilation can reduce peak live activations but may increase transient workspaces; report both max allocated (live set) and max reserved (allocator cache) when profiling.
4. Optimizer memory depends on choice: SGD+momentum uses 12 bytes/parameter vs AdamWs 16; parameter-side savings are larger for ResNet-50 due to its smaller parameter count.

# 6 Verification Recipe (PyTorch)

Measure live vs reserved memory on a single forward+backward to validate environment-specific peaks:

```
import torch
def measure(model, x):
    model.train()
    torch.cuda.empty_cache()
    torch.cuda.reset_peak_memory_stats()
    out = model(x).sum()
    out.backward()
    alloc = torch.cuda.max_memory_allocated() / (1024**3)
    reserv = torch.cuda.max_memory_reserved() / (1024**3)
    return alloc, reserv
# Example shapes:
# ResNet-50: x = torch.randn(64, 3, 224, 224, device='cuda')
# ViT-B/16: x = torch.randn(64, 3, 224, 224, device='cuda')
```

Report `max_memory_allocated` (peak live set). `max_memory_reserved` includes the allocator cache and can exceed true usage.

# 7 Summary

- **ResNet-50** (FP32, AdamW, 224×224, $B = 64$): $\sim$ 3.03 GiB ex-workspace; $\sim$ 3.333.63 GiB with headroom.
- **DeiT-Base / ViT-B/16**:
  - Naïve MHA: $\sim 6.55$ GiB ex-workspace; $\sim 7.217.86$ GiB with headroom.
  - SDPA/Flash: $\sim 5.22$ GiB ex-workspace; $\sim 5.746.27$ GiB with headroom.
- The transformers higher memory arises from larger parameter count and higher $O(BSH)$ activation volume per layer; the quadratic $O(B\,h\,S^2)$ component is implementation-dependent and avoidable with modern attention kernels.

# 8 Appendix: Constants and Conversions

- FP32: 4 bytes/element
- 1 MiB = $1{,}048{,}576$ bytes, 1 GiB = $1{,}073{,}741{,}824$ bytes
- Batch = 64; Image = $224 \times 224$; ViT sequence $S = 197$; ViT hidden $H = 768$; Heads $h = 12$; Layers $L = 12$.