FLOPs and MACs Calculation for DistilBERT-Base

We will calculate the floating-point operations (FLOPs) and multiply-accumulate operations (MACs) for a single forward pass of the DistilBERT-Base model (6 layers, 12 attention heads, hidden size 768) on the following input sentence:

"I love deep learning. It's very interesting."

We assume:

- WordPiece tokenization
- FP16 precision, where each multiply–add pair is considered a single MAC (multiply–accumulate).
- A hidden size $H = 768$
- 12 attention heads (heads = 12), so each head has dimensionality $d_k = H/\text{heads} = 64$
- An FFN inner dimension $I = 3072$ (i.e. 4×768)
- 6 Transformer layers (DistilBERT typically has half the layers of BERT).

1. Tokenization

Tokenize the sentence "I love deep learning. It's very interesting." using WordPiece, adding `[CLS]` and `[SEP]`:

$$[CLS],\ i,\ love,\ deep,\ learning,\ .,\ it,\ \text{“}\#\#'s\text{”},\ very,\ interesting,\ .,\ [SEP]$$

We obtain $N = 12$ tokens.

2. Embeddings

DistilBERT learns a token embedding and a positional embedding of size 768. We add these two vectors elementwise for each token:

$$\text{Embeddings} = \underbrace{\text{Lookup(token\_id)} + \text{Lookup(position)}}_{\text{768-dim each}}.$$

- Lookups are not counted as FLOPs, as they are essentially indexing.
- Adding two 768-dim vectors for each token costs 768 additions per token.
- For N tokens, that is $N \times 768$ FLOPs.

Hence, for $N = 12$:

$$\text{Embedding\_FLOPs} = 12 \times 768 = 9216 \quad (0 \text{ MACs}).$$

3. Multi-Head Self-Attention (per layer)

Each layer has a multi-head self-attention (MHA) mechanism with 12 heads. The hidden size is 768, and each head operates on vectors of length $d_k = 64$. We break it down:

Q, K, V projections. For each token (length 768), we compute Query (Q), Key (K), and Value (V) vectors by multiplying by weight matrices of shape 768×768 (one for Q, one for K, one for V):

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V$$

Each is a matrix multiply for dimension $(N \times 768) \times (768 \times 768)$. For one projection, multiply and add counts are each about $N \times 768 \times 768$. For three projections (Q,K,V):

$$\text{Mult}_{QKV} = 3 \times N \times 768 \times 768, \quad \text{Add}_{QKV} \approx 3 \times N \times 768 \times (768 - 1).$$

The MAC count equals the number of multiplications. For $N = 12$:

$$\text{MACs}_{QKV} = 3 \times 12 \times 768 \times 768 = 21{,}233{,}664.$$

Total FLOPs is multiplications plus additions, i.e. $\approx 42{,}439{,}680$.

Attention scores $Q \cdot K^\top$. We form AttnScores $= QK^\top$ for each head. Each head has $N \times 64$ Q and $N \times 64$ K, so $N \times N$ dot products of length 64 per head. Each dot product has 64 multiplies and 63 adds:

$$\text{Mult}_{QK} = \text{heads} \times N \times N \times 64, \quad \text{Add}_{QK} \approx \text{heads} \times N \times N \times 63.$$

Hence MACs equals the multiply count. For $N = 12$, 12 heads:

$$\text{MACs}_{QK} = 12 \times 12^2 \times 64 = 110{,}592.$$

Total FLOPs $\approx$ 219,456.

Scaling, Softmax. We scale by $1/\sqrt{64}$, then apply softmax along each row. Softmax includes exponentiations, sums, and divisions. The total here is on the order of a few thousand FLOPs, negligible compared to the large matrix multiplies.

Applying softmax$(QK^\top)$ to V. We multiply an $N \times N$ attention matrix by $N \times d_k$ to produce $N \times d_k$. Per head, that costs $N \times N \times d_k$ multiplies plus $(N \times N \times (d_k - 1))$ adds. For 12 heads, that is the same scale as $QK^\top$:

$$\text{MACs}_{Attn \times V} = \text{heads} \times N \times N \times 64 = 110{,}592.$$

Total FLOPs $\approx$ 219,456 again.

Output projection. The 12 heads are concatenated into $N \times 768$, then multiplied by a 768×768 output weight. This is again $N \times 768 \times 768$ multiplies and similar adds. For $N = 12$:

$$\text{MACs}_O = 12 \times 768 \times 768 = 7{,}077{,}888,$$

FLOPs $\approx$ 14,146,560.

Summarizing MHA per layer:

$$\text{MHA\_FLOPs} \approx 56.99\text{M},$$
$$\text{MHA\_MACs} \approx 28.53\text{M}.$$

4. Feed-Forward Network (per layer)

Each layer has a two-layer FFN with dimensions $768 \to 3072 \to 768$ and a GELU nonlinearity in between.

First linear: 768×3072. For each token, we multiply a 768-dim vector by a 768×3072 matrix, so 768×3072 multiplies per token plus additions. For N tokens,

$$\text{MACs}_{\text{ffn1}} = N \times 768 \times 3072.$$

For $N = 12$, that is 28,311,552 MACs. FLOPs is about twice that (mult + add) $\approx$ 56.6M.

GELU activation. GELU is more complex than ReLU. Approximate 4 FLOPs per element. For $N \times 3072$ elements, $12 \times 3072 \times 4 = 147{,}456$ FLOPs, negligible compared to the matrix multiplies.

Second linear: 3072×768. Similarly, $N \times 3072 \times 768$ multiplies and similar adds. Another 28,311,552 MACs. $\approx$ 56.6M FLOPs.

Hence total FFN per layer:

$$\text{FFN\_FLOPs} \approx 113.2\text{M}, \quad \text{FFN\_MACs} \approx 56.62\text{M}.$$

5. Layer Normalization (per layer)

We apply LayerNorm twice per layer (after MHA, after FFN). For each token (768-dim):

1. Compute mean (768 adds, 1 div).
2. Compute variance (768 sub, 768 mul, sum, 1 div).
3. Normalize each value, multiply by gamma, add beta (768 sub, 768 mul, 768 mul, 768 add, 1 sqrt, 1 div).

Total ≈ 6147 FLOPs per token per LayerNorm. For $N = 12$:

$$\text{FLOPs}_{\text{LayerNorm}} \approx 6147 \times 12 = 73{,}764 \quad \text{(per LN)}.$$

Two LN calls per layer: 147,528 FLOPs per layer. Negligible MACs.

6. Residual Connections (per layer)

Each sub-layer output is added to the original input (dimension 768) per token. Two residual connections (after MHA, after FFN):

$$\text{FLOPs}_{\text{residual}} = 2 \times N \times 768.$$

For $N = 12$, $2 \times 12 \times 768 = 18{,}432$ FLOPs per layer.

7. Total (per layer) and for 6 Layers

Combining each component for a single layer:

$$\begin{aligned}
\text{MHA_FLOPs} &\approx 56.99\text{M}, \\
\text{FFN_FLOPs} &\approx 113.2\text{M}, \\
\text{LayerNorm} &\approx 0.147\text{M}, \\
\text{Residual} &\approx 0.018\text{M}, \\
\Rightarrow \text{Per-layer FLOPs} &\approx 170.54\text{M}.
\end{aligned}$$

$$\begin{aligned}
\text{MHA_MACs} &\approx 28.53\text{M}, \\
\text{FFN_MACs} &\approx 56.62\text{M}, \\
\text{LayerNorm} &\approx 0, \\
\text{Residual} &\approx 0, \\
\Rightarrow \text{Per-layer MACs} &\approx 85.15\text{M}.
\end{aligned}$$

DistilBERT has 6 layers, so multiply by 6:

$$\begin{aligned}
\text{6-layer FLOPs} &\approx 6 \times 170.54\text{M} \approx 1.02324 \times 10^9, \\
\text{6-layer MACs} &\approx 6 \times 85.15\text{M} \approx 5.109 \times 10^8.
\end{aligned}$$

Finally, add the Embedding cost (9216 FLOPs), negligible. No MACs in embedding. So total is:

$$\boxed{1.023 \times 10^9 \text{ FLOPs} \quad \text{and} \quad 5.11 \times 10^8 \text{ MACs}.}$$