

All the Transformer Math You Need to Know

Part 4 of [How To Scale Your Model \(Part 3: Sharding | Part 5: Training\)](#)

Here we'll do a quick review of the Transformer architecture, specifically how to calculate FLOPs, bytes, and other quantities of interest.

AUTHORS

[Jacob Austin](#)
[Sholto Douglas](#)
[Roy Frostig](#)
[Anselm Levskaya](#)
[Charlie Chen](#)
[Sharad Vikram](#)

[Federico Lebron](#)
[Peter Choy](#)
[Vinay Ramasesh](#)
[Albert Webster](#)
[Reiner Pope*](#)

AFFILIATION

Google DeepMind

PUBLISHED

Feb. 4, 2025

Counting Dots

Let's start with vectors x, y and matrices A, B of the following shapes:

array	shape
x	[P]
y	[P]
A	[N P]
B	[P M]

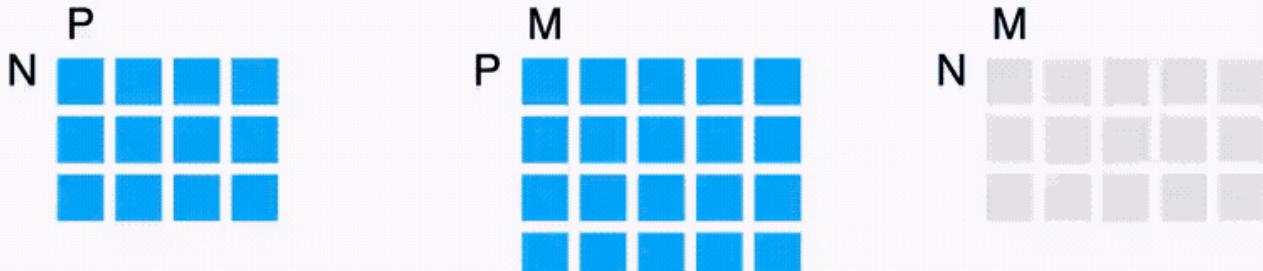
- A dot product of $x \cdot y$ requires P adds and multiplies, or $2P$ floating-point operations total.
 - A matrix-vector product Ax does N dot-products along the rows of A , for $2NP$ FLOPs.
 - A matrix-matrix product AB does a matrix-vector product for each of the M columns of B , for $2NPM$ FLOPs total.
 - In general, if we have two higher dimensional arrays C and D , where some dimensions are CONTRACTING and some are BATCHING. (e.g. $C[GHIJKL], D[GHMNL]$) then the FLOPs cost of this contraction is two times the product of all of the C and D dimensions where the batch and contraction dimensions are only counted once, (e.g. $2GHIJMNKL$). Note that a dimension is only batching if it occurs in both multiplicands. (Note also that the factor of 2 won't apply if there are no contracting dimensions and this is just an elementwise product.)
- (More precisely, P mult + $P-1$ add
N.M - (2?) ignore
contract

Operation	FLOPs	Data (memory reads)
$x \cdot y$	$2P$	$2P$
Ax	$2NP$	$NP + P$
AB	$2NPM$	$NP + PM$
$[c_0, \dots, c_N] \cdot [d_0, \dots, d_N]$	$2 \prod c_i \times \prod_{\substack{d_j \notin \text{BATCH} \\ d_j \in \text{CONTRACT}}} d_j$	$\prod c_i + \prod d_j$

Make note of the fact that for a matrix-matrix multiply, the compute scales cubically $O(N^3)$ while the data transfer only scales quadratically $O(N^2)$ - this means that as we scale up our matmul size, it becomes easier to hit the compute-saturated limit. This is extremely unusual, and explains in large part why we use architectures dominated by matrix multiplication - they're amenable to being scaled!

when A, B
[N, N] [N, N]





Forward and reverse FLOPs

During training, we don't particularly care about the result of a given matrix multiply; we really care about its derivative. That means we do significantly more FLOPs during backpropagation.

If we imagine \mathbf{B} is just one matrix in a larger network and \mathbf{A} are our input activations with $\mathbf{C} = \mathbf{A} \mathbf{B}$, the derivative of the loss \mathbf{L} with respect to \mathbf{B} is given by the chain rule:

$$\frac{\partial \mathbf{L}}{\partial \mathbf{B}} = \frac{\partial \mathbf{L}}{\partial \mathbf{C}} \frac{\partial \mathbf{C}}{\partial \mathbf{B}} = \mathbf{A}^T \left(\frac{\partial \mathbf{L}}{\partial \mathbf{C}} \right)$$

which is an outer product and requires $2NPM$ FLOPs to compute (since it contracts over the N dimension). Likewise, the derivative of the loss with respect to \mathbf{A} is

$$\frac{\partial \mathbf{L}}{\partial \mathbf{A}} = \frac{\partial \mathbf{L}}{\partial \mathbf{C}} \frac{\partial \mathbf{C}}{\partial \mathbf{A}} = \left(\frac{\partial \mathbf{L}}{\partial \mathbf{C}} \right) \mathbf{B}^T$$

the derivative of input activation is used to compute parameter gradient

is again $2NPM$ FLOPs since $d\mathbf{L}/d\mathbf{C}$ is a (co-)vector of size $[N, M]$. While this quantity isn't the derivative w.r.t. a parameter, it's used to compute derivatives for previous layers of the network (e.g. just as $d\mathbf{L}/d\mathbf{C}$ is used to compute $d\mathbf{L}/d\mathbf{B}$ above).

Important takeaway! Forward pass gradient

Adding these up, we see that during training, we have a total of $6NPM$ FLOPs, compared to $2NPM$ during inference: $2NPM$ in the forward pass, $4NPM$ in the backward pass. Since PM is the number of parameters in the matrix, this is the simplest form of the famous $6 * \text{num parameters} * \text{num tokens}$ approximation of Transformer FLOPs during training: each token requires $6 * \text{num parameters}$ FLOPs. We'll show a more correct derivation below.

Transformer Accounting

Transformers are the future. Well, they're the present at least. Maybe a few years ago, they were one of many architectures. But today, it's worth knowing pretty much every detail of the architecture. We won't reintroduce the architecture but [this blog](#) and the [original Transformer paper](#) may be helpful references.

Here's a basic diagram of the Transformer decoder architecture:



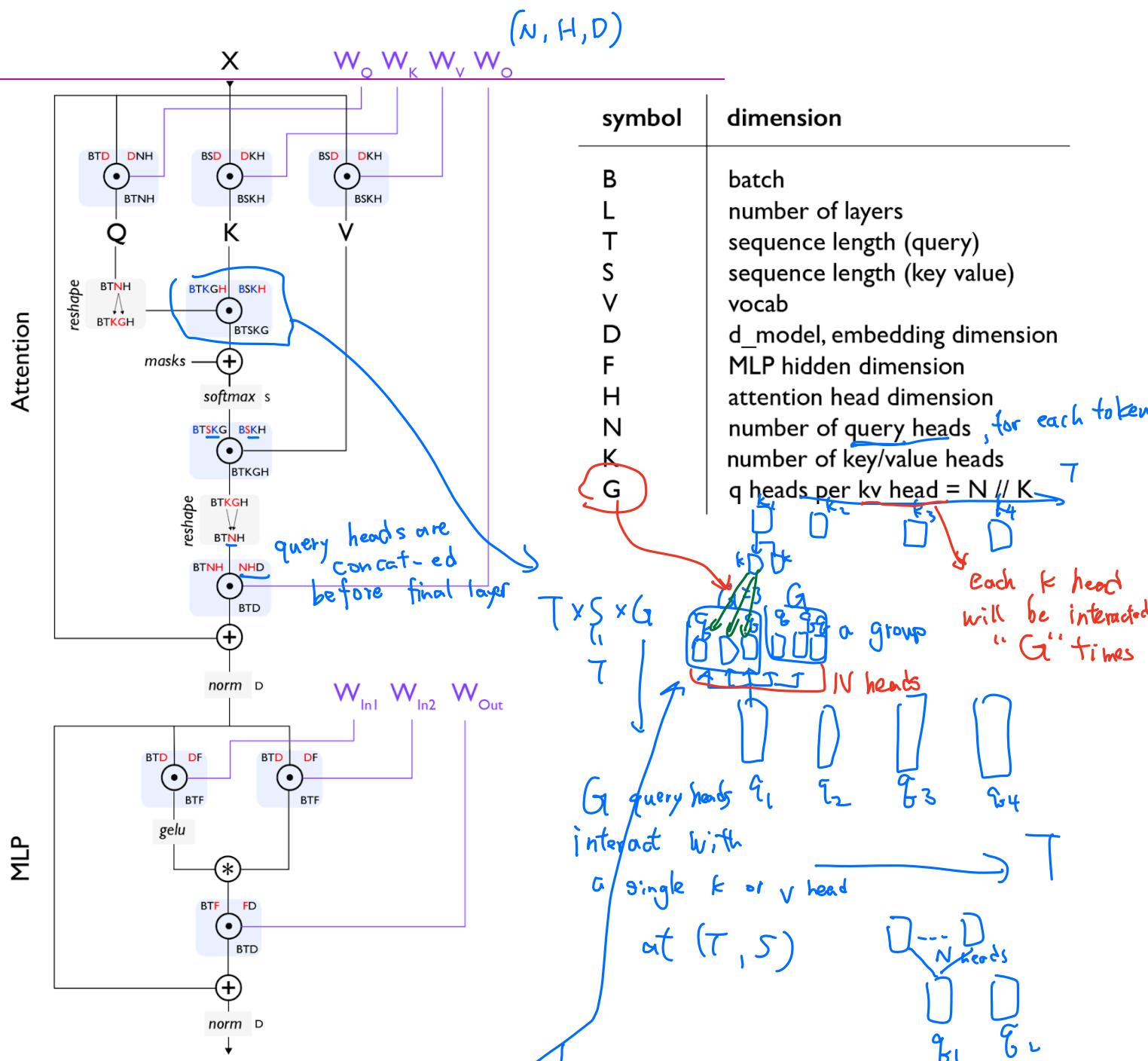


Figure: this diagram shows one layer of a standard Transformer and flows from top-to-bottom. We use a single-letter convention to describe the shapes and layouts of arrays in a Transformer, again showing contracting dimensions in red, and batched dimensions in blue. In a given operation, the input shape is given on top-left and the parameter shape is given on the top-right, with the resulting shape below, e.g. BTD is the input shape for the gating einsum and DF is the weight shape.

Note [gating einsum]: The diagram above uses a “[gating einsums](#)” [1] where we split the up-projection matrix into two matrices ($W_{\text{In}1}$ and $W_{\text{In}2}$ above) whose outputs are elementwise multiplied as a kind of “gating function”. Not all LMs use this, so you will sometimes see a single W_{In} matrix and a total MLP parameter count of 2DF instead of 3DF. Typically in this case, D and F will be scaled up to keep the parameter count the same as the 3 matrix case. With that said, some form of gating einsum is used by LLAMA, DeepSeek, and many other models.

Note 2 [MHA attention]: With self-attention, T and S are the same but for cross-attention they may be different. With vanilla Multi-Head Attention (MHA), N and K are the same while for [Multi-Query Attention](#) (MQA) [2] $K=1$ and for [Grouped MQA](#) (GMQA) [3] K merely has to divide N.

when $N=K \Rightarrow G=1$ key only generates one single head

For the below we're going to compute per-layer FLOPs to avoid having to stick factors of L everywhere.

MLPs

The MLPs of a Transformer typically consist of 2 input matmuls that are element-wise combined and a single output matmul:

operation	train FLOPs	params
$A[B, T, D] \cdot W_{in1}[D, F]$	<u>6BTDF</u>	DF
$A[B, T, D] \cdot W_{in2}[D, F]$	<u>6BTDF</u>	DF
$\sigma(A_{in1})[B, T, F] * A_{in2}[B, T, F]$	$O(BTF)$	<u>ignore</u>
$A[B, T, F] \cdot W_{out}[F, D]$	<u>6BTDF</u>	DF
		$\approx 18BTDF$
		$3DF$

w

Attention

For the generic grouped-query attention case with different **Q** and **KV** head numbers, let us assume equal head dimension **H** for **Q,K,V** projections, and estimate the cost of the **QKVO** matmuls:

operation	train FLOPs	params
$A[B, T, D] \cdot W_Q[D, N, H]$	<u>6BTDNH</u>	<u>DNH</u>
$A[B, T, D] \cdot W_K[D, K, H]$	<u>6BTDKH</u>	DKH
$A[B, T, D] \cdot W_V[D, K, H]$	<u>6BTDKH</u>	DKH
$A[B, T, N, H] \cdot W_O[N, H, D]$	<u>6BTDNH</u>	DNH
		$\rightarrow N = K \cdot G$ when vanilla multihead attention
	$12BTD(N + K)H$	$2D(N + K)H$

The dot-product attention operation is more subtle, effectively being a $TH \cdot HS$ matmul batched over the B, K dimensions, a softmax, and a $TS \cdot SH$ matmul again batched over the B, K dimensions. We highlight the batched dims in blue:

operation	train FLOPs
$Q[B, T, K, G, H] \cdot K[B, S, K, H]$	<u>6BTSKGH</u> = <u>6BTSNH</u>
softmax _S $L[B, T, S, K, G]$	$O(BTSKG) = O(BTSN)$
$S[B, T, S, K, G] \cdot V[B, S, K, H]$	<u>6BTSKGH</u> = <u>6BTSNH</u>
	$\approx 12BTSNH = 12BT^2NH$

Note [causal masking]: Most recent transformers use a causal mask as opposed to full bidirectional attention. In this case the useful FLOPs of the dot product operations are reduced by a factor of 1/2. To achieve this reduction in practice we need to make use of an attention kernel, rather than a naive einsum.

Other Operations

There are several other operations happening in a Transformer. Layernorms are comparatively cheap and can be ignored for first-order cost estimates. There is also the final enormous (though not per-layer) unembedding matrix multiply.

\hookrightarrow recover index to vocabulary

operation	train FLOPs	params
layernorm _D $A[B, T, D]$	$O(BTD)$	D
$A[B, T, D] \cdot W_{unembed}[D, V]$	<u>6BTDV</u>	DV

General rule of thumb for Transformer FLOPs

If we neglect the cost of dot-product attention for shorter-context training, then the total FLOPs across all layers is



$$(18BTDF + 12BTD(N+K)H)L = 6 * BT * [3DF + 2D(N+K)H]L$$

$$= 6 * \text{num tokens} * \text{parameter count}$$

parameter-related

Leading to a famous rule of thumb for estimating dense Transformer FLOP count, ignoring the attention FLOPs. (Unembedding is another simple matmul with $6BTDV$ FLOPs and DV params, and follows the same rule of thumb.)

Fractional cost of attention with context length

If we do account for dot-product attention above and assume $F = 4D$, $D = NH$ (as is typical) and $N = K$:

$$\frac{\text{attention FLOPs}}{\text{matmul FLOPs}} = \frac{12BT^2NH}{18BTDF + 24BTDNH} = \frac{12BT^2D}{4 * 18BTD^2 + 24BTD^2} = \frac{12BT^2D}{96BTD^2} = \frac{T}{8D}$$

Original paper
multi-head
attention

So the takeaway is that dot-product attention FLOPs only become dominant during training once $T > 8D$. For $D \sim 8k$, this would be $\sim 64k$ tokens. This makes some sense, since it means as the MLP size increases, the attention FLOPs become less critical. For large models, the quadratic cost of attention is not actually a huge obstacle to longer context training. However, for smaller models, even e.g. Gemma-27B, $D=4608$ which means attention becomes dominant around 32k sequence lengths. Flash Attention also helps alleviate the cost of long-context, which we discuss briefly in [Appendix A](#).

Miscellaneous Math

Sparsity and Mixture-of-Experts

Careless

We'd be remiss not to briefly discuss Mixture of Experts (MoE) models [4], which replace the single dense MLP blocks in a standard Transformer with a set of independent MLPs that can be dynamically routed between. To a first approximation, an MoE is just a normal dense model with E MLP blocks per layer, instead of just one. Each token activates k of these experts, typically $k \ll E$. The ratio E/k is called the sparsity and is usually between 8 and 64 (e.g.

[DeepSeek v3](#) has effectively $k = 8, E = 256$). This increases the parameter count by $O(E)$, while multiplying the total number of activated parameters per token by k , compared with the dense version.

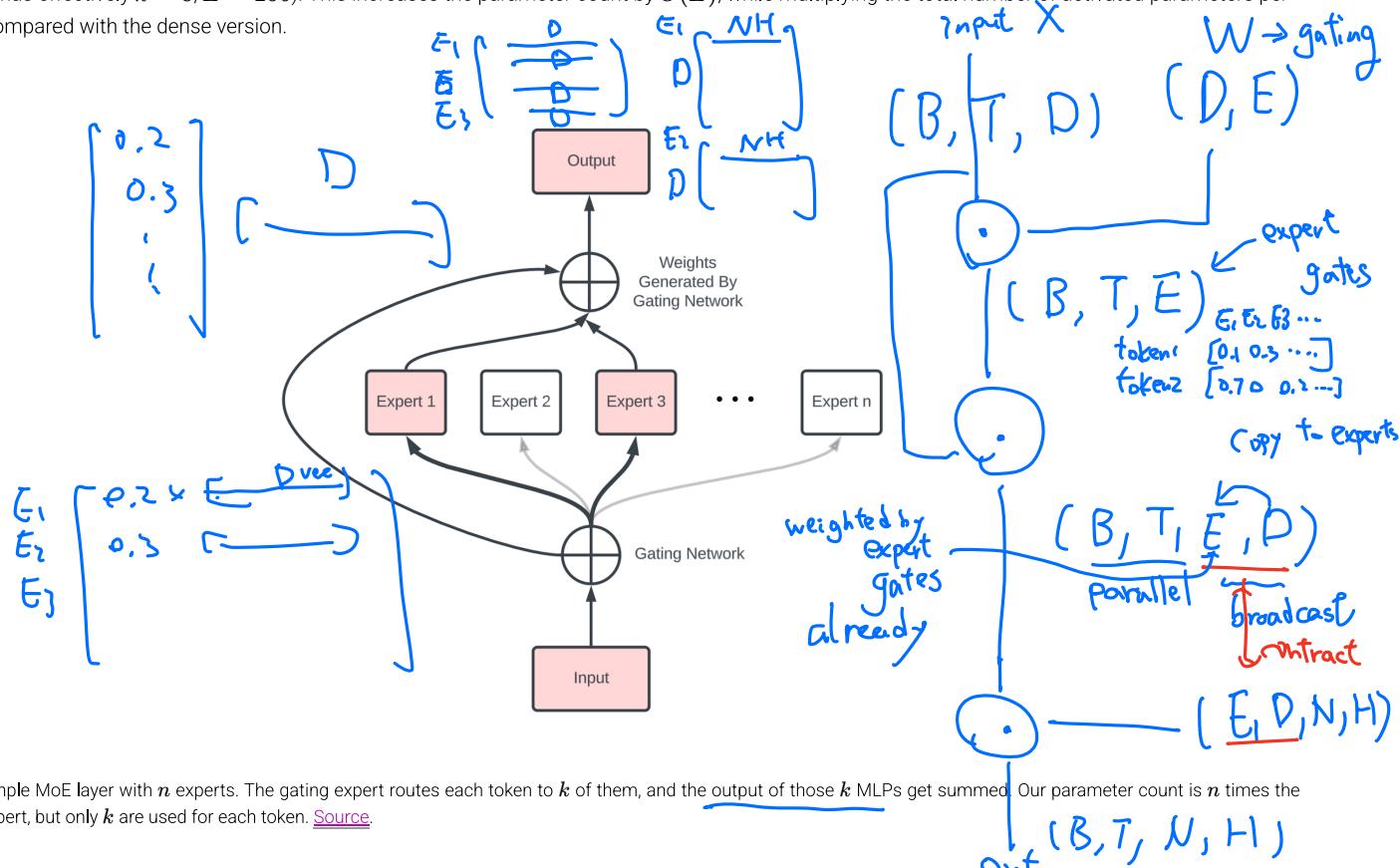


Figure: an example MoE layer with n experts. The gating expert routes each token to k of them, and the output of those k MLPs get summed. Our parameter count is n times the size of each expert, but only k are used for each token. [Source](#).

Compared to a dense model, an MoE introduces new comms, primarily two AllToAlls (one before and one after the MoE block) that route tokens to the correct expert and brings them back to their home device.¹ However as we saw in the previous section, the cost of each AllToAll is only 1/4 that of a comparable AllGather along a single axis (for a bidirectional ring).

Gradient checkpointing

Backpropagation as an algorithm trades memory for compute. Instead of a backward pass requiring $O(n_{\text{layers}}^2)$ FLOPs, it requires $O(n_{\text{layers}})$ memory, saving all intermediate activations generated during the forward pass. While this is better than quadratic compute, it's incredibly expensive memory-wise: a

model with $B * T = 4M$ (4M total tokens per batch), $L=64$, and $D=8192$ that avoids all unnecessary backward pass compute would have to save roughly $2 * 20 * \frac{B * T * D * L}{2} = 84TB$ of activations in bfloat16. 20 comes from (roughly) counting every intermediate node in the Transformer diagram above, since e.g.

$$40 \times 4M \times 64 \times 8192 = 84 \times 10^{12} \approx 84TB \text{ number of activations}$$

$$f(x) = \exp(g(x))$$

$$\frac{df}{dx} = \exp(g(x)) \cdot \frac{dg}{dx}$$

so to avoid recomputing we need to save $g(x)$ and $\exp(g(x))$ from the forward pass. To avoid saving this much memory, we can choose to only save some fraction of the intermediate activations. Here are a few strategies we use.

- Block remat:** only save the input to each layer. This is the most aggressive method we use and only saves 1 checkpoint per layer, meaning we'd only save 4.2TB in the example above. This forces us to repeat essentially all forward pass FLOPs in the backward pass, meaning we increase our FLOPs from $6ND$ to roughly $8ND$.
- Big matmuls only:** another simple policy is to only save the outputs of large matmuls. This lets us avoid recomputing any large matmuls during the backward pass, but still makes us recompute other activation functions and parts of attention. This reduces 20 per layer to closer to 7 per layer.

This by no means comprehensive. When using JAX, these are typically controlled by `jax.remat` / `jax.checkpoint` (you can read more [here](#)).

Key-Value (KV) caching

As we'll see in [Section 7](#), LLM inference has two key parts, prefill and generation.

- Prefill** processes a long prompt and saves its attention activations in a Key-Value Cache (KV Cache) for use in generation, specifically the key-value projections in the attention block.
- Generation** batches several of these KV caches together and samples tokens from each of them.

Each KV cache is then effectively an array of size $[2, S, L, K, H]$ where the 2 accounts for the keys and values. This is quite large! The total size of the Key-Value cache in int8 is $2SLKH$. For a moderately-sized model with 8k context length, 64 layers, and $KH = NH = D = 8192$, this is $2 \cdot 8192 \cdot 64 \cdot 8192 = 8\text{GiB}$. You can see why we would want to use GMQA with $K \ll N$.

KH

\sim multiple query heads looking at the same query head

What Should You Take Away from this Section?

- The overall parameters and FLOPs of a Transformer are fairly easy to calculate, and are summarized here, assuming MHA (with batch size B , vocab size V , a sequence of length T , $D=d_{\text{model}}$, and $F=d_{\text{ff}}$):

Component	Params per layer	Training FLOPs per layer
MLP	3DF	18BTDF
Attention	4DNH	$24BDNH + 12BT^2NH$
Other	D	BT
Vocab	DV (total, not per-layer)	12BDV

per layer

$2DV$, one for embedding, one for mapping back to vocab

- The parameter count of the MLP block dominates the total parameter count and the MLP block also dominates the FLOPs budget as long as the sequence length $T < 8D$.
- The total FLOPs budget during training is well approximated by $6 \cdot \text{num_params} \cdot \text{num_tokens}$ for reasonable context lengths.
- During inference, our KV caches are roughly $2 \cdot S \cdot L \cdot K \cdot H$ per cache (where K is the number of KV heads), although architectural modifications can often reduce this.

A Few Problems to Work

Question 1: How many parameters does a model with $D = 4096$, $F = 4 \cdot D$, $V = 32,000$, and $L = 64$ have? What fraction of these are attention parameters? How large are our KV caches per token? You can assume $N \cdot H = D$ and multi-head attention with int8 KVs.

► Click here for the answer.

Question 2: How many total FLOPs are required to perform $A[B_x, D_y] *_D W[D_y, F]$ on `{'X': 4, 'Y': 8, 'Z': 4}`. How many FLOPs are performed by each TPU?

► Click here for the answer.



Question 3: How many FLOPs are involved in performing $A[I, J, K, L] * B[I, J, M, N, O] \rightarrow C[K, L, M, N, O]$?

► Click here for the answer.

Question 4: What is the arithmetic intensity of self-attention (ignoring the Q/K/V/O projections)? Give the answer as a function of the Q and KV lengths T and S. At what context length is attention FLOPs-bound? Given the HBM bandwidth of our TPUs, plot the effective relative cost of attention to the FFW block as the context length grows.

► Click here for the answer.

Question 5: At what sequence length are self-attention FLOPs equal to the QKVO projection FLOPs?

► Click here for the answer.

↳ this question seems similar to Question 5
Should $T: 2D$ ratio

Question 6: Say we only save the output of each of the 7 main matmuls in a Transformer layer during our forward pass (Q, K, V, O + the three FFW matrices). How many extra FLOPs do we need to "rematerialize" during the backwards pass?

► Click here for the answer.

Question 7: DeepSeek v3 says it was trained for 2.79M H800 hours on 14.8T tokens ([source](#)). Given that it has 37B activated parameters, roughly what hardware utilization did they achieve? Hint: note that they used FP8 FLOPs without structured sparsity.

► Click here for the answer.

Question 8: Mixture of Experts (MoE) models have E copies of a standard dense MLP block, and each token activates k of these experts. What batch size in tokens is required to be compute-bound for an MoE with weights in int8 on TPU v5e? For DeepSeek, which has 256 (routed) experts and $k = 8$, what is this number?

► Click here for the answer.

That's it for Part 4! For Part 5 (about scaling Transformer training), [click here!](#)

Appendix

Appendix A: How does Flash Attention work?

The traditional objection to scaling Transformers to very long context is that the attention FLOPs and memory usage scale quadratically with context length. While it's true that the attention QK product has shape $[B, T, S, N]$ where B is the batch size, S and T are the Q and K sequence dims, and N is the number of heads, this claim comes with some serious caveats:

1. As we noted earlier, even though this is quadratic, the attention FLOPs only dominate when $S > 8 \cdot D$, and especially during training the memory of a single attention matrix is small compared to all of the weights and activation checkpoints living in memory, especially when sharded.
2. We don't need to materialize the full attention matrix in order to compute attention! We can compute local sums and maxes and avoid ever materializing more than a small chunk of the array. While the total FLOPs is still quadratic, we drastically reduce memory pressure.

This second observation was first made by [Rabe et al. 2021](#) and later in the [Flash Attention paper](#) (Dao et al. 2022). The basic idea is to compute the attention in chunks of K/V, where we compute the local softmax and some auxiliary statistics, then pass them onto the next chunk which combines them with its local chunk. Specifically, we compute

1. **M:** The running max of $q \cdot k$ over the sequence dimension
2. **O:** The running full attention softmax over the sequence dimension
3. **L:** The running denominator $\sum_i (q \cdot k_i - \text{running max})$

i.e.

$$\sum_i \frac{\exp(q \cdot k_i - \max_j q \cdot k_j)}{\sum_l \exp(q \cdot k_l - \max_j q \cdot k_j)} \cdot v_i$$

With these, we can compute the new max, the new running sum, and the new output with only a constant amount of memory. To give a sketchy description of how this works, attention is roughly this operation:



Q vector, K vector, V vector

$$\text{Attn}(Q, K, V) = \sum_i \frac{\exp(Q \cdot K_i - \max_j Q \cdot K_j) V_i}{\sum_l \exp(Q \cdot K_l - \max_j Q \cdot K_j)}$$

numerical stability

The max is subtracted for numerical stability and can be added without affecting the outcome since $\sum_i \exp(a_i + b) = \exp(b) \sum \exp(a)$. Looking just at the denominator above, if we imagine having two contiguous chunks of key vectors, K^1 and K^2 and we compute the local softmax sums L^1 and L^2 for each

$$L^1 = \sum_i \exp(Q \cdot K_i^1 - \max_j Q \cdot K_j^1)$$

$$L^2 = \sum_i \exp(Q \cdot K_i^2 - \max_j Q \cdot K_j^2)$$

$$\begin{aligned} O^1 &= \sum_i \frac{\exp(Q \cdot K_i^1 - \max_j Q \cdot K_j^1) V_i}{\sum_l \exp(Q \cdot K_l^1 - \max_j Q \cdot K_j^1)} \\ O^2 &= \sum_i \frac{\exp(Q \cdot K_i^2 - \max_j Q \cdot K_j^2) V_i}{\sum_l \exp(Q \cdot K_l^2 - \max_j Q \cdot K_j^2)} \end{aligned}$$

Then we can combine these into the full softmax sum for these two chunks together by using

$$L^{\text{combined}} = \exp(M^1 - \max(M^1, M^2)) \cdot L^1 + \exp(M^2 - \max(M^1, M^2)) \cdot L^2$$

where

$$M^1 = \max_j Q \cdot K_j^1 \text{ and } M^2 = \max_j Q \cdot K_j^2$$

$$O^{\text{combined}} = \frac{\exp(M^1) \cdot O^1 + \exp(M^2) \cdot O^2}{\exp(M^{\text{combined}}) + \exp(M^{\text{combined}})}$$

This can be done for the full softmax as well, giving us a way of accumulating arbitrarily large softmax sums. Here's the full algorithm from the Flash Attention paper.

Algorithm 1 FLASHATTENTION

Require: Matrices $Q, K, V \in \mathbb{R}^{N \times d}$ in HBM, on-chip SRAM of size M .

- 1: Set block sizes $B_c = \lceil \frac{M}{4d} \rceil$, $B_r = \min(\lceil \frac{M}{4d} \rceil, d)$.
- 2: Initialize $O = (0)_{N \times d} \in \mathbb{R}^{N \times d}$, $\ell = (0)_N \in \mathbb{R}^N$, $m = (-\infty)_N \in \mathbb{R}^N$ in HBM.
- 3: Divide Q into $T_r = \lceil \frac{N}{B_r} \rceil$ blocks Q_1, \dots, Q_{T_r} of size $B_r \times d$ each, and divide K, V in to $T_c = \lceil \frac{N}{B_c} \rceil$ blocks K_1, \dots, K_{T_c} and V_1, \dots, V_{T_c} , of size $B_c \times d$ each.
- 4: Divide O into T_r blocks O_i, \dots, O_{T_r} of size $B_r \times d$ each, divide ℓ into T_r blocks $\ell_i, \dots, \ell_{T_r}$ of size B_r each, divide m into T_r blocks m_1, \dots, m_{T_r} of size B_r each.
- 5: **for** $1 \leq j \leq T_c$ **do**
- 6: Load K_j, V_j from HBM to on-chip SRAM.
- 7: **for** $1 \leq i \leq T_r$ **do**
- 8: Load Q_i, O_i, ℓ_i, m_i from HBM to on-chip SRAM.
- 9: On chip, compute $S_{ij} = Q_i K_j^T \in \mathbb{R}^{B_r \times B_c}$.
- 10: On chip, compute $\tilde{m}_{ij} = \text{rowmax}(S_{ij}) \in \mathbb{R}^{B_r}$, $\tilde{P}_{ij} = \exp(S_{ij} - \tilde{m}_{ij}) \in \mathbb{R}^{B_r \times B_c}$ (pointwise), $\tilde{\ell}_{ij} = \text{rowsum}(\tilde{P}_{ij}) \in \mathbb{R}^{B_r}$.
- 11: On chip, compute $m_i^{\text{new}} = \max(m_i, \tilde{m}_{ij}) \in \mathbb{R}^{B_r}$, $\ell_i^{\text{new}} = e^{m_i - m_i^{\text{new}}} \ell_i + e^{\tilde{m}_{ij} - m_i^{\text{new}}} \tilde{\ell}_{ij} \in \mathbb{R}^{B_r}$.
- 12: Write $O_i \leftarrow \text{diag}(\ell_i^{\text{new}})^{-1} (\text{diag}(\ell_i) e^{m_i - m_i^{\text{new}}} O_i + e^{\tilde{m}_{ij} - m_i^{\text{new}}} \tilde{P}_{ij} V_j)$ to HBM.
- 13: Write $\ell_i \leftarrow \ell_i^{\text{new}}$, $m_i \leftarrow m_i^{\text{new}}$ to HBM.
- 14: **end for**
- 15: **end for**
- 16: Return O .

From a hardware standpoint, this lets us fit our chunk of Q into VMEM (what the algorithm above calls on-chip SRAM) so we only have to load the KV chunks on each iteration, increasing the arithmetic intensity. We can also keep the running statistics in VMEM.

One last subtle point worth emphasizing is an attention softmax property that's used to make the Flash VJP (reverse mode derivative) calculation practical for training. If we define an intermediate softmax array as:

$$S_{ij} = \frac{e^{\tau q_i \cdot k_j}}{\sum_l e^{\tau q_i \cdot k_l}}$$

In attention, we obtain dS from reverse-mode dO and V arrays:

$$S_{\text{matrix}} \cdot V_{\text{matrix}} = O_{\text{matrix}}$$

$$\frac{\partial L}{\partial S} = \frac{\partial L}{\partial Q} \frac{\partial Q}{\partial K}$$

During the backpropagation of this gradient to Q and K

Note: derivation below

$$dS_{ij} = dO_{id} \cdot_d V_{jd} = \sum_d dO_{id} V_{jd}$$

contract

$$S_{ij} = \frac{\exp(q_i k_j)}{\sum_k \exp(q_i k_k)}$$

We exploit an identity that allows us to exchange a contraction along the large key length dimension with a local contraction along the feature depth dimension.

$$\begin{aligned} S_{ij} \cdot_j dS_{ij} &= \sum_j \frac{e^{\tau q_i \cdot k_j}}{\sum_k e^{\tau q_i \cdot k_k}} \sum_d dO_{id} V_{jd} \\ &= \sum_d dO_{id} \sum_j \frac{e^{\tau q_i \cdot k_j}}{\sum_k e^{\tau q_i \cdot k_k}} V_{jd} \\ &= \sum_d dO_{id} O_{id} \\ &= dO_{id} \cdot_d O_{id} \end{aligned}$$

This replacement is crucial for being able to implement a sequence-block local calculation for the VJP, and enables further clever sharding schemes like ring attention.

Takeaway: The input gradient to softmax can be computed from $\frac{\partial L}{\partial Q}$ row and O's row, where d does not depend on seq length.

Otherwise, the need to go through all time dimension would have forced us to non-local computation

Footnotes

1. Technically, this only happens if we are data or sequence sharded along the same axis as our experts. [2]

References

1. GLU Variants Improve Transformer
Shazeer, N., 2020. arXiv [cs.LG].
2. Fast Transformer decoding: One write-head is all you need
Shazeer, N., 2019. arXiv [cs.NE].
3. GQA: Training generalized multi-query transformer models from multi-head checkpoints
Ainslie, J., Lee-Thorp, J., de Jong, M., Zemlyanskiy, Y., Lebrón, F. and Shanghai, S., 2023. arXiv [cs.CL].
4. Outrageously large neural networks: The Sparsely-Gated Mixture-of-experts layer
Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G. and Dean, J., 2017. arXiv [cs.LG].

Miscellaneous

*Work done at Google DeepMind, now at MatX.

Citation

For attribution in academic contexts, please cite this work as:

Austin et al., "How to Scale Your Model", Google DeepMind, online, 2025.

or as a BibTeX entry:

```
@article{scaling-book,
  title = {How to Scale Your Model},
  author = {Austin, Jacob and Douglas, Sholto and Frostig, Roy and Levskaya, Anselm and Chen, Charlie and Vikram, Sharad and Lebron, Federico and Choy, Peter and Ramasesh, Vinay and Webson, Albert and Pope, Reiner},
  publisher = {Google DeepMind},
  howpublished = {Online},
  note = {Retrieved from https://jax-ml.github.io/scaling-book/},
  year = {2025}
}
```



14 comments · 22 replies - powered by giscus

Oldest

Newest

© Copyright 2026 . Powered by [Jekyll](#) with [al-folio](#) theme. Hosted by [GitHub Pages](#).Softmax gradient : (T, T)

$$S = \text{softmax}(X) \quad (\text{over columns})$$

impact

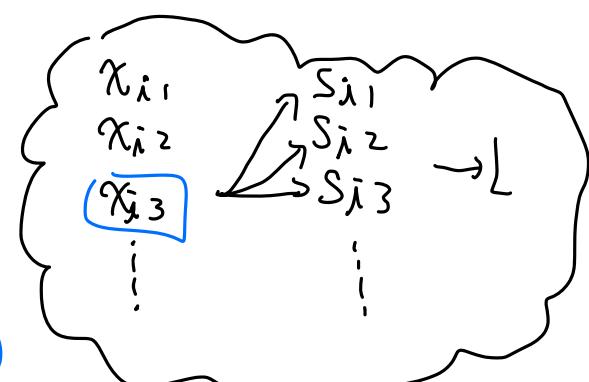
$$S = \begin{bmatrix} \text{---} \\ \text{---} \\ \text{---} \end{bmatrix} \quad X = \begin{bmatrix} \text{---} \\ \text{---} \\ \text{---} \end{bmatrix}$$

$$S_{ij} = \frac{\exp(X_{ij})}{\sum_l \exp(X_{il})}$$

If we have : $\frac{\partial L}{\partial S_{ij}}$ already , compute $\frac{\partial L}{\partial X_{ij}}$

$$\frac{\partial L}{\partial X_{ij}} = \sum_k \frac{\partial L}{\partial S_{ik}} \cdot \frac{\partial S_{ik}}{\partial X_{ij}}$$

chain rule



$$\frac{\partial S_{ij}}{\partial X_{ij}} = \frac{\partial}{\partial X_{ij}} \frac{\exp(X_{ij})}{\exp(X_{ij}) + \sum_l \exp(X_{il})}$$

f(x) *g(x)*

$$= \frac{\exp(X_{ij}) \frac{\partial X_{ij}}{\partial X_{ij}} g(x)}{\left(\sum_l \exp(X_{il}) \right)^2} - \frac{\exp(X_{ij}) \left(\sum_l \exp(X_{il}) \right) f'(x)}{\left(\sum_l \exp(X_{il}) \right)^2}$$

(recap: $\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$)

$$= \frac{\exp(X_{ij})}{\sum_l \exp(X_{il})} - \left(\frac{\exp(X_{ij})}{\sum_l \exp(X_{il})} \right)^2$$

$$= \frac{\exp(X_{ij})}{\sum_l \exp(X_{il})} - \left(\frac{\exp(X_{ij})}{\sum_l \exp(X_{il})} \right)^2 = S_{ij} - S_{ij}^2$$



$$\frac{\partial S_{ik}}{\partial x_{ij}} \text{ (where } k \neq j) = \boxed{\exp(x_{ik})} \cdot \frac{-1}{\left(\sum_{l} \exp(x_{il})\right)^2} \cdot \exp(x_{ij})$$

Constant to x_{ij}

$$= \boxed{-S_{ik} \cdot S_{ij}}$$

Putting together:

$$\frac{\partial L}{\partial x_{ij}} = \sum_k \frac{\partial L}{\partial S_{ik}} \cdot \frac{\partial S_{ik}}{\partial x_{ij}}$$

$$= \frac{\partial L}{\partial S_{ij}} \cdot \boxed{\frac{\partial S_{ij}}{\partial x_{ij}}} + \sum_{k \neq j} \frac{\partial L}{\partial S_{ik}} \cdot \frac{\partial S_{ik}}{\partial x_{ij}}$$

$$= \frac{\partial L}{\partial S_{ij}} \cdot \boxed{(S_{ij} - S_{ij}^2)} + \sum_{k \neq j} \frac{\partial L}{\partial S_{ik}} \cdot \boxed{(-S_{ik} \cdot S_{ij})}$$

$$= S_{ij} \cdot \frac{\partial L}{\partial S_{ij}} + \frac{\partial L}{\partial S_{ij}} (-S_{ij}^2) + \sum_{k \neq j} \frac{\partial L}{\partial S_{ik}} (-S_{ik} \cdot S_{ij})$$

$\underbrace{\quad}_{\text{merge}}$

$$= S_{ij} \cdot \frac{\partial L}{\partial S_{ij}} - \sum_k \frac{\partial L}{\partial S_{ik}} S_{ik} \cdot \boxed{S_{ij}}$$

\uparrow

$$= S_{ij} \cdot \left[\frac{\partial L}{\partial S_{ij}} - \sum_k \frac{\partial L}{\partial S_{ik}} S_{ik} \right] \#$$

\downarrow sequence dimension

Focusing on the : $\left[\sum_k \frac{\partial L}{\partial s_{ik}} \cdot s_{ik} \right]$ part

$$\sum_k \frac{\partial L}{\partial s_{ik}} \cdot s_{ik} \quad \frac{\partial L}{\partial s} = \frac{\partial L}{\partial o} V^T \rightarrow V \in \mathbb{R}^{\text{seq-length}}$$

$$= \sum_k \underline{s_{ik}} \cdot \left(\sum_d \frac{\partial L}{\partial o_{id}} \cdot V_{kd} \right)$$

$$= \sum_k \sum_d \frac{\partial L}{\partial o_{id}} \cdot s_{ik} \cdot V_{kd}$$

\swarrow swap

$$= \sum_d \sum_k \frac{\partial L}{\partial o_{id}} \cdot s_{ik} V_{kd}$$

\swarrow take this out

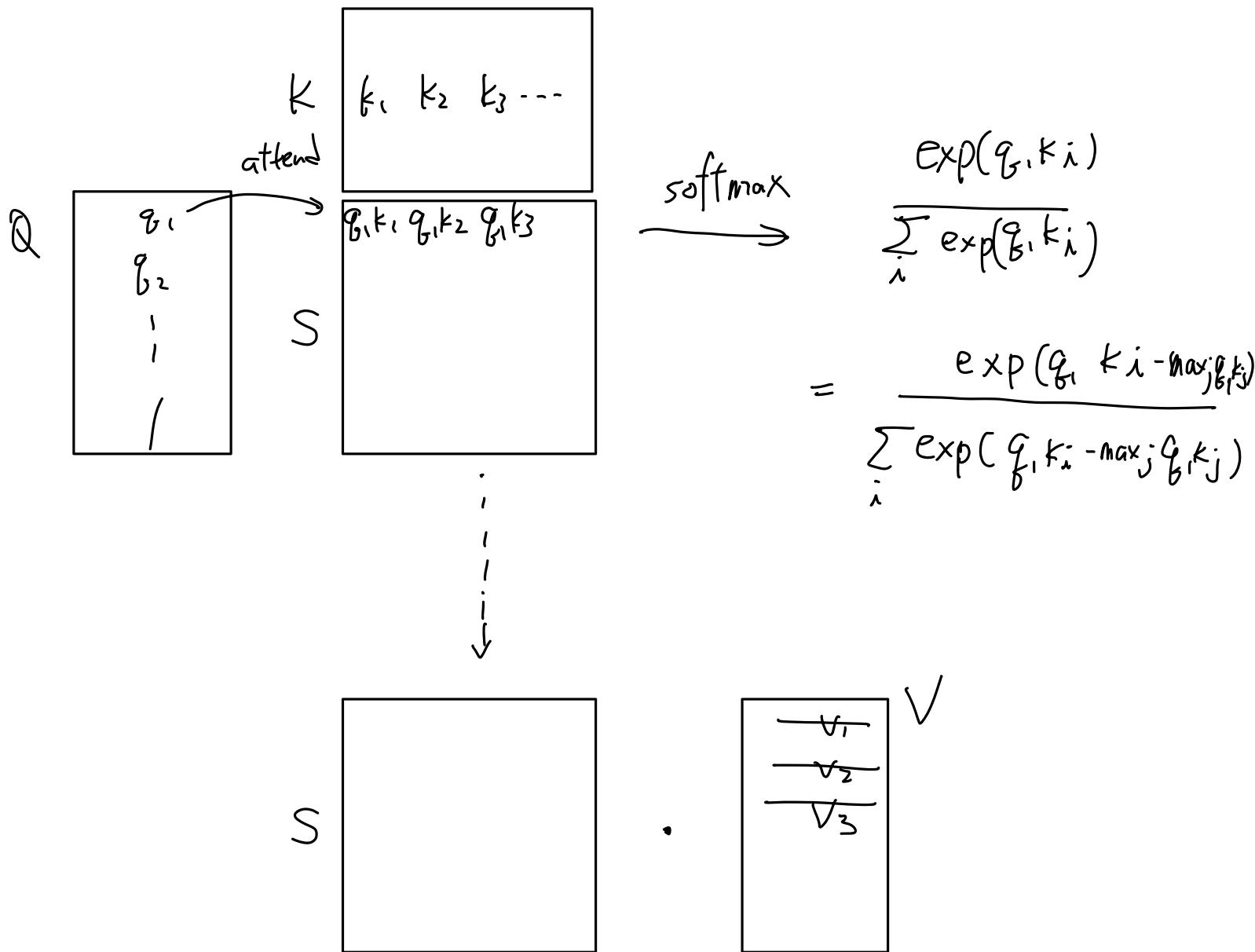
$$= \sum_d \frac{\partial L}{\partial o_{id}} \left(\sum_k s_{ik} V_{kd} \right)$$

$$= \sum_d \frac{\partial L}{\partial o_{id}} \left(\underline{s_{ik} \cdot V_{kd}} \right)$$

$$= \sum_d \frac{\partial L}{\partial o_{id}} \cdot O_{id}$$

by definition, this is O_{id}

Flash attention recap:



Problems :

Question 1: How many parameters does a model with $D = 4096$, $F = 4 \cdot D$, $V = 32,000$, and $L = 64$ have? What fraction of these are attention parameters? How large are our KV caches per token? You can assume $N \cdot H = D$ and multi-head attention with int8 KVs.

$$\text{MLPs} = 3DF \times L = 3 \cdot 4096 \cdot (4 \cdot 4096) \times 64 = 12 \times 2^{12} \times 2^{12} \times 2^6$$

$$\text{Attention} = 4DNH \times L = 4 \cdot (4096 \cdot 4096) \times 64 = 2^{20} \times 2^{12} \times 2^{12} \times 2^6$$

$$\text{Other} = D \times L = 4096 \times 64 = 2^{12} \times 2^6 \times 2^{18}$$

$$+) \text{Vocab} = \underbrace{2DV}_{\sim \text{embed} + \text{unembed}} = 2 \times 4096 \times 32,000 = 2 \times 2^{12} \times 32000$$

$$\text{Total} = 12TiB + 4TiB + \text{parameters} \\ 0.25 GiB + 0.26 GiB$$

$$\text{Attention fraction} = \frac{4TiB}{(6TiB)} = 25\%$$

$$\text{Per token } K \text{ size} = \sum_{k=1}^K NH = NH = D \\ = 4096 \text{ bytes}$$

$$\text{All layers } K \text{ size} = 64 \times 4096 \text{ bytes}$$

$$\text{All layers } K \text{ and } V \text{ cache size} = 2 \times 64 \times 4096 \\ = 2 \times 2^6 \times 2^{12} = 2^{19} \text{ bytes} \\ \approx 0.5 MiB$$

Question 2: How many total FLOPs are required to perform $A[B_X, D_Y] *_D W[D_Y, F]$ on $\{X: 4, Y: 8, Z: 4\}$. How many FLOPs are performed by each TPU?

$$A[B_X, D_Y] \cdot W[D_Y, F] \xrightarrow{\text{multiply}} O[B_X, F]\{U_Y\}$$

$$O[B_X, F]\{U_Y\} \xrightarrow{\text{AllReduce} \quad (\text{FLOPs in add})} O[B_X, F] \xrightarrow{\text{unreduced}}$$

$$\text{Multiply} = 2 \frac{B}{X} \cdot \frac{P}{Y}, F = \frac{2 BDF}{4 \times 8} = \frac{BDF}{16}$$

$$\text{Add from AllReduce} = \text{ReduceScatter} + \cancel{\text{AllGather}}_{\text{no flops}}$$

= FLOPs from ReduceScatter

$$= \frac{B}{|X|} \times \frac{F}{|Y|} \times \left(\frac{|X|}{2} \right) \times 2 \leftarrow \begin{array}{l} \text{(the number of hops)} \\ \text{bidirectional add} \end{array}$$

$$= \frac{BF}{4}$$

$$\text{Sum} = \frac{BDF}{16} + \frac{BF}{4}$$

Question 3: How many FLOPs are involved in performing $A[I, J, \underline{K}, L] * B[I, J, \underline{M}, N, O] \rightarrow C[K, L, M, N, O]$?



$$\# \text{ of FLOPs} = 2 \sum_{I,J} KLMNO$$

Question 4: What is the arithmetic intensity of self-attention (ignoring the Q/K/V/O projections)? Give the answer as a function of the Q and KV lengths T and S. At what context length is attention FLOPs-bound? Given the HBM bandwidth of our TPUs, plot the effective relative cost of attention to the FFW block as the context length grows.

$$Q = (B, T, N, H) \xrightarrow{\text{reshape}} (B, T, K, G, H)$$

$$K, V = (B, S, K, H) \xrightarrow{T=S} (B, T, K, H)$$

batch dimension

contract

Assume : $T = S$ when doing self-attention
(prefill)

a) FLOPs for attention matrix (B, T, T, K, G)

$$= 2 \cdot B \cdot K \cdot T \cdot T \cdot G \cdot H$$

$$= 2 \cdot B \cdot T^2 \cdot K \cdot G \cdot H$$

b) FLOPs for softmax = element-wise add row sum

$$= O(C \cdot B \cdot T^2 \cdot K \cdot G)$$

⑥ FLOPs for attention and V matrix

$$= [B, T, \underbrace{T}_{\text{Contract over context}}, K, G] * [B, \underbrace{T, K, H}_{\text{Contract over context}}]$$
$$\rightarrow (B, T, K, G, I-H)$$

$$= 2 \cdot B T \cdot T \cdot K \cdot G \cdot H$$

$$= 2 B T^2 K G H$$

Bytes to load = $\begin{cases} \text{Load } Q, K + \text{Write attention} \\ \text{Load attention} + \text{Load } V + \text{Write attended results} \end{cases}$

assume bfloat16

B SKH

$$\textcircled{1} = B T K G H + 2 B T T K G$$

flash attention
avoids
this

$$\textcircled{2} = 2 B T T K G + 2 B T K G H$$

Arithmetic intensity

$$= \frac{① + ② + ③}{① + ②}$$

$$= \frac{2BT^2KGH + O(BT^2KG) + 2BT^2KGH}{4BTKGH + 2BT^2KG + 2BT^2KG + 2BTKGH}$$

$$= \frac{4BT^2KGH + O(BT^2KG)}{6BTKGH + 4BT^2KG}$$

$$\approx \frac{2}{3}T = O(T) \text{ (prefill)}$$

$$\text{If TPU Arithmetic intensity} = \frac{1.97 \times 10^{14}}{8.2 \times 10^{11}}$$

$$\approx 240$$

$$T > 240 \Rightarrow \text{compute bound}$$

After prefill, generation $T = 1$,

so AI becomes = $\frac{O(BTSKGH)}{O(BTKGH) + O(BSKH)}$

= $\frac{O(BSKGH)}{O(BKGH) + O(BSKI)}$

= $\frac{O(BSKGH)}{O(BKH \cdot (G+S))}$

= $\frac{O(SG)}{O(G+S)}$

S usually dominates

so = $O(G)$ which is small

Question 5: At what sequence length are self-attention FLOPs equal to the QKVO projection FLOPs?

$$\text{FLOPs attention} = 4BTSKGH$$

$$\text{FLOPs QKVO projections} = \underbrace{Q}_{+} = 2BTDKGH$$

$$\underbrace{K}_{+} = 2BSDKH$$

$$\underbrace{V}_{+} = 2BSDKH$$

$$\underbrace{O}_{\substack{= 2BTN \\ \parallel \\ KG}} = 2BTN - 1D$$

$$= 4BTDKGH + 4BSDKH$$

Assume prefill $T=S$ and $\frac{N}{G}=k$
 $(G=1)$

$$= 4BTDKGH + 4BTDKH$$

$$= 8BTDKH$$

Solve

$$4\cancel{BTTKGH} = 8\cancel{BTDKH}$$

$$\Rightarrow T = 2D$$

If $D = 4096$, then $T = 8192$ so
usually for small contexts, projections

are more expensive.

Question 6: Say we only save the output of each of the 7 main matmuls in a Transformer layer during our forward pass (Q, K, V, O + the three FFW matrices). How many extra FLOPs do we need to "rematerialize" during the backwards pass?

If we only store Q, K, V, O and $O_{W_{in_1}}, O_{W_{in_2}}$, $O_{W_{out}}$,

to compute $\frac{\partial L}{\partial W_Q}, \frac{\partial L}{\partial W_K}, \frac{\partial L}{\partial W_V}, \frac{\partial L}{\partial W_O}$,

$\frac{\partial L}{\partial W_{in_1}}, \frac{\partial L}{\partial W_{in_2}}, \frac{\partial L}{\partial W_{out}}$

$\frac{\partial L}{\partial W_Q}, \frac{\partial L}{\partial W_K}, \frac{\partial L}{\partial W_V}$: need $X \Rightarrow$ assume already there?

$\frac{\partial L}{\partial W_O} =$ heed input to W_O
 $= \text{softmax}(QK) \vee = O(4BTSKGH)$
when $T=S \Rightarrow 4BT^2KGH$

$\frac{\partial L}{\partial W_{in_1}}, \frac{\partial L}{\partial W_{in_2}} =$ need to recompute input

$= \text{norm}((\text{the output of } W_O) + X)$

$\frac{\partial L}{\partial W_{in_1}} = O(BTD)$

$\frac{\partial L}{\partial W_{in_2}} =$ need to recompute input

$= \text{gelu}(\text{output of } W_{in_1}) \times (\text{output of } W_{in_2})_{\text{element-wise}}$

$\Rightarrow O(BTF)$

Question 7: DeepSeek v3 says it was trained for 2.79M H800 hours on 14.8T tokens ([source](#)). Given that it has 37B activated parameters, roughly what hardware utilization did they achieve? Hint: note that they used FP8 FLOPs without structured sparsity.

$$\text{Approximate FLOPs} = 6 \cdot 37B \times 14.8T \text{ tokens}$$

$$= 6 \times 37 \times 10^9 \times 14.8 \times 10^{12}$$

$$= 3285.6 \times 10^{21}$$

$$\approx 3.3 \times 10^{24} \text{ FLOPs}$$

$$\text{FLOPs per sec} = \frac{3.3 \times 10^{24}}{2.79 \times 10^6 \times 60 \times 60}$$

$$= \frac{3.3 \times 10^{24}}{100.44 \times 10^8}$$

$$= 0.032 \times 10^{16}$$

$$= 3.2 \times 10^{14} \text{ FLOP/sec}$$

FP8 FLOPs H800 per sec

$$= 3000 T \text{ FLOPs / sec} / 2 \leftarrow \text{answer says half}$$

$$= 3000 \times 10^{12} / \text{sec} = 3 \times 10^{15} = \text{peak}$$

$$\text{FLOP utilization} = \frac{3.2 \times 10^{14}}{3 \times 10^{15}} \times 2 = \cancel{10\%} \approx 20\%$$

Question 8: Mixture of Experts (MoE) models have E copies of a standard dense MLP block, and each token activates k of these experts. What batch size in tokens is required to be compute-bound for an MoE with weights in int8 on TPU v5e? For DeepSeek, which has 256 (routed) experts and $k = 8$, what is this number?

$$\text{MoE: } [B, T, D] \times (E, D, F)$$

Each token is routed to k experts

In average, each expert gets $\frac{BT}{k}$ tokens, in average

So the Compute on each expert is

$$(E, \frac{BT}{k}, D) \times (E, D, F) = (E, \frac{BT}{k} D, F)$$

↓ ↓
batch Contract

$$\text{FLOPs} = 2 E \times \frac{BT}{k} \times D F = \frac{2EBTDF}{k}$$

$$\text{ReadWrite} = \frac{EBTD}{k} + EDF + \frac{EBTF}{k}$$

$$= \frac{EDT(D+F)}{k} + EDF + \frac{BT(F+k)}{k}$$

$$\text{Arithmetic Intensity}_{\text{TPUv5e}} = \frac{3.94 \times 10^{14}}{8.1 \times 10^{11}} = 0.48 \times 10^3 = 480$$

^{1.97} ^{0.24}
use bf16?

$$\text{Arithmetic Intensity}_{\text{MoE}} = \frac{2BTDFk}{(EDF + BT(F+k))}$$

Compute - bound when

$$\frac{2BTDF_k}{EDF + BT_k(D+F)} > 240$$

$$\Rightarrow 2BTDF_k > 240 EDF + \\ 240 BT_k D + \\ 240 BT_k F$$

/DFK

$$\Rightarrow 2BT > 240 E/k + \frac{240 BT}{F} + \frac{240 BT}{D}$$

$$\Rightarrow 2BT - \frac{240 BT}{F} - \frac{240 BT}{D} > 240 E/k$$

$$\Rightarrow \left(2 - \underbrace{\frac{240}{F}}_{\text{should be very small}} - \underbrace{\frac{240}{D}}_{\text{should be very small}} \right) BT > 240 E/k$$

$$\Rightarrow \underbrace{2BT}_{\# \text{ of tokens}} > 240 E/k$$

$$\Rightarrow BT > 120 \frac{E}{K}$$

$$\text{when } E/K = \frac{256}{8} = 32$$

$$\Rightarrow BT > 120 \times 32 \\ = 3840$$

So MoE is only compute-bound when there are at least $O(3840)$ tokens