Decoder-Only Transformers

Notes on various aspects of Decoder-Only Transformers. Conventions are in App. A.

Contents

Part I
Architecture

1 Decoder-Only Fundamentals

The Transformers architecture [1], which dominates Natural Language Processing (NLP) as of July 2023, is a relatively simple architecture. There are various flavors and variants of Tranformers, but focus here on the decoder-only versions which underlie the GPT models [2–4].

The full decoder-only architecture can be seen in Fig. 1. The parameters which define the network can be found in App. A.

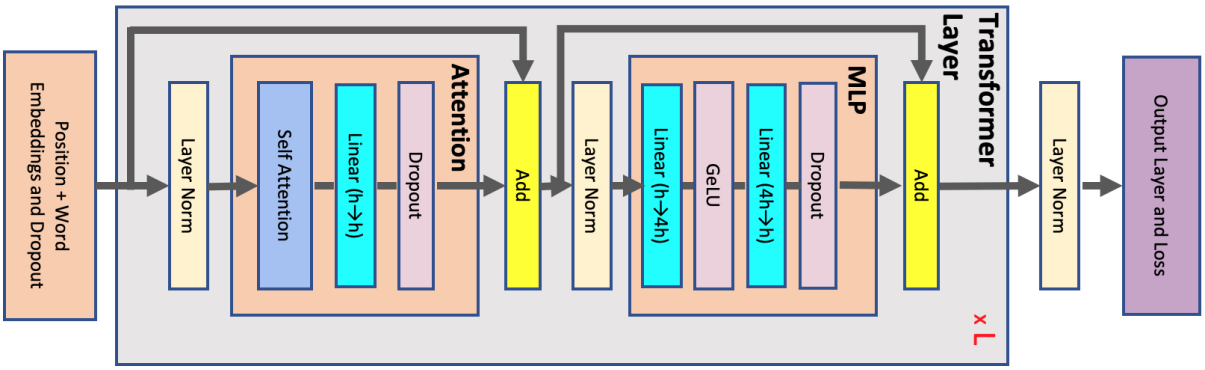

Figure 1. The full transformers architecture. Diagram taken from [5]

At a high level, decoder-only transformers take in an ordered series of word-like objects, called tokens, and are trained to predict the next token in the sequence. Given some initial text, transformers can be used to give a prediction for the likelihood of any possible continuation of that text. An outline of the mechanics[1]:

1. Raw text is **tokenized** and turned into a series of integers[2] whose values lie in `range(V)`, with V the vocabulary size.

2. The tokenized text is chunked and turned into `(B, S)`-shaped (batch size and sequence length, respectively) integer tensors, x_{bs}.

3. The **embedding layer** converts the integer tensors into continuous representations of shape `(B, S, D)`, z_{bsd}, with D the size of the hidden dimension. **Positional encodings** have also been added to the tensor at this stage to help the architecture understand the relative ordering of the text.

4. The z_{bsd} tensors pass through a series of transformer blocks, each of which has two primary components:

[1] This describes the vanilla architecture; almost every component is modified in the available variants.
[2] There are about 1.3 tokens per word, on average.

(a) In the **attention** sub-block, components of z_{bsd} at different positions (s-values) interact with each other, resulting in another `(B, S, D)`-shaped tensor, z'_{bsd}.

(b) In the **MLP** block, each position in z'_{bsd} is processed independently and in parallel by a two-layer feed-forward network, resulting once more in a `(B, S, D)`-shaped tensor.

Importantly, there are **residual connections** around each of these[3] (the arrows in Fig. 1), meaning that the output of each block is added back to its original input.

5. Finally, we convert the `(B, S, D)`-shaped tensors to `(B, S, V)`-shaped ones, y_{bsv}. This is the role of the **language model head** (which is often just the embedding layer used in an inverse manner.)

6. The y_{bsv} predict what the next token will be, i.e. x_{bs+1}, having seen the **context** of the first s tokens in the sequence. Specifically, removing the batch index for simplicity, a `Softmax` of y_{sv} gives the conditional probability $p_{sv} = P(t_{s+1}|t_s \dots t_0)$ for the indicated series of tokens. Because of the chain rule of probability, these individual probabilities can be combined to form the probability that any sequence of tokens follows a given initial seed[4].

Each batch (the b-index) is processed independently. We omitted `LayerNorm` and `Dropout` layers above, as well as the causal mask; these will be covered below as we step through the architecture in more detail.

1.1 Embedding Layer and Positional Encodings

The **embedding** layer is just a simple look up table: each of the `range(V)` indices in the vocabulary is mapped to a D-dimensional vector via a large `(V, D)`-shaped table/matrix. This layer maps $x_{bs} \longrightarrow z_{bsd}$. In `torch`, this is an `nn.Embedding(V, D)` instance.

To each item in a batch, we add identical **positional encodings** to the vectors above with the goal of adding fixed, position-dependent correlations in the sequence dimension which will hopefully make it easier for the architecture to pick up on the relative positions of the inputs [5] This layer maps $z_{bsd} \leftarrow z_{bsd} + p_{sd}$, with p_{sd} the positional encoding tensor.

The above components require $(V + S)D \approx VD$ parameters per model.

1.2 Layer Norm

The original transformers paper [1] put `LayerNorm` instances after the **attention** and **MLP** blocks, but now it is common [6] to put them before these blocks[6].

[3] This gives rise to the concept of the **residual stream** which each transformer block reads from and writes back to repeatedly.

[4] In more detail, these probabilities are created by products: $P(t_{s+n} \dots t_{s+1}|t_s \dots t_0) = P(t_{s+n}|t_{s+n-1} \dots t_s \dots t_0) \times \dots \times P(t_{s+1}|t_s \dots t_0)$.

[5] Positional encodings and the causal mask are the only components in the vanilla transformers architecture which carry weights with a dimension of size S; i.e. they are the only parts that have explicit sequence-length dependence. A related though experiment: you can convince yourself that if the inputs z_{bsd} were just random noise, the transformers architecture would not be able to predict the s-index of each such input in the absence of positional encodings.

[6] Which makes intuitive sense for the purposes of stabilizing the matrix multiplications in the blocks

The `LayerNorm` operations acts over the hidden dimension (since this is the dimension the subsequent `Linear` instances act on). Spelling it out, given the input tensor z_{bsd} whose mean and variance over the d-index are μ_{bs} and σ_{bs}, respectively, the `LayerNorm` output is

$$z_{bsd} \leftarrow \left(\frac{z_{bsd} - \mu_{bs}}{\sigma_{bs}} \right) \times \gamma_d + \beta_d \equiv \text{LayerNorm}_{\,d}\, z_{bsd} \tag{1.1}$$

where γ_d, β_d are the trainable scale and bias parameters. In `torch`, this is a `nn.LayerNorm(D)` instance. Since there are two `LayerNorm` instances in each transformer block, these components require $2D$ parameters per layer.

We will continue discussing `LayerNorm` instances in what follows in order to adhere to the usual construction and to discuss methods like sequence-parallelism in their original form (see Sec. 10.4), but note: the data-independent `LayerNorm` transformations due to γ_d, β_d are completely redundant when immediately followed by a `Linear` layer, since both act linearly on their inputs and `Linear` is already the most general data-independent linear transformation. Explicitly, the γ_d, β_d parameters can be absorbed into the `Linear` parameters:

$$(x_{bsd}\gamma_d + \beta_d)\, W_{dd'} + b_{d'} = x_{bsd} W'_{dd'} + b'_{d'} \ , \quad W'_{dd'} \equiv \gamma_d W_{dd'} \ , \quad b'_{d'} \equiv b_{d'} + \beta_d W_{dd'} \ , \tag{1.2}$$

for arbitrary x_{bsd}. That is, these transformations can be equivalently performed by the weight matrix and bias (if included) in `Linear` layer[7].

1.3 Causal Attention

Causal attention is the most complex layer. It features A sets of weight matrices[8] $Q_{dea}, K_{dea}, V_{dea}$ where $a \in \{0, \ldots, A-1\}$ and $e \in \{0, \ldots, D/A\}$, where D is assumed perfectly divisible by A. From these, we form three different vectors:

$$q_{bsea} = z_{bsd} Q_{dea} \ , \quad k_{bsea} = z_{bsd} K_{dea} \ , \quad v_{bsea} = z_{bsd} V_{dea} \tag{1.3}$$

These are the **query, key, and value** tensors, respectively [9].

Using the above tensors, we will then build up an **attention map** $w_{bss'a}$ which corresponds to how much attention the token at position s pays to the token at position s'. Because we have the goal of predicting the next token in the sequence, we need these weights to be causal: the final prediction y_{bsv} should only have access to information propagated from positions $x_{bs'v}$ with $s' \leq s$. This corresponds to the condition that $w_{bss'a} = 0$ if $s' > s$. The entire causal Transformers architecture as a whole obeys this condition: the outputs $z_{bsd} = \texttt{CausalTransformer}(x_{bs'd'})$ only depend on those inputs $x_{bs'd'}$ with $s' \leq s$.

[7] Note the importance of data-independence here: the data-dependent mean and standard deviation terms cannot be similarly absorbed. Also, because the usual training algorithms are not invariant under parameter redefinitions, the above unfortunately does not imply that removing the learnable `LayerNorm` parameters (`elementwise_affine=False` in `torch`) will have no effect on training dynamics. γ_d, β_d can shoved into the `Linear` layer's parameters as a small inference-time optimization, though.

[8] There are also bias terms, but we will often neglect to write them explicitly or account for their (negligible) parameter count.

[9] There are of course many variants of the architecture and one variant which is popular in Summer 2023 is multi-query attention [7] in which all heads share *the same* key and value vectors and only the query changes across heads, as this greatly reduces inference costs. See Sec. 2.2.

These weights come from `Softmax`-ed attention scores, which are just a normalized dot-product over the hidden dimension:

$$w_{bss'da} = \texttt{Softmax}_{s'}\left(m_{ss'} + \frac{q_{bse}k_{bs'ea}}{\sqrt{D/A}}\right) \quad , \quad \text{s.t.} \quad \sum_{s'} w_{bdss'a} = 1 \tag{1.4}$$

The tensor $m_{ss'}$ is the causal mask which zeroes out the relevant attention map components above

$$m_{ss'} = \begin{cases} 0 & s \leq s' \\ -\infty & = s > s' \end{cases} \quad ,$$

forcing $w_{bss'da} = 0$ for $s > s'$. In other words, the causal mask ensures that a given tensor, say z_{bsd}, only has dependence on other tensors whose sequence index, say s', obeys $s' \leq s$. This is crucial for inference-time optimizations, in particular the use of the **kv-cache** in which key-value pairs do not need to be re-computed.

The $\sqrt{D/A}$ normalization is motivated by demanding that the variance of the `Softmax` argument be 1 at initialization, assuming that other components have been configured so that that the query and key components are i.i.d. from a Gaussian normal distribution [10].

The weights above are then passed through a dropout layer and used to re-weigh the **value** vectors and form the tensors

$$y_{bsea} = \text{Drop}\left(w_{bdss'a}\right) v_{bs'ea} \tag{1.5}$$

and these `(B, S, D/A, A)`-shaped tensors are then concatenated along the e-direction to re-form a `(B, S, D)`-shaped tensor u_{bsd}

$$u_{bsd} = y_{bs(ea)} \tag{1.6}$$

in `einops`-like notation for concatenation. Finally, another weight matrix $O_{d'd}$ and dropout layer transform the output once again to get the final output

$$z_{bsd} = \text{Drop}\left(u_{bsd'}O_{d'd}\right) \ . \tag{1.7}$$

For completeness, the entire operation in condensed notation with indices left implicit is:

$$z \leftarrow \text{Drop}\left(\text{Concat}\left(\text{Drop}\left(\texttt{Softmax}\left(\frac{(z\cdot Q_a)\cdot(z\cdot K_a)}{\sqrt{D/A}}\right)\right)\cdot z\cdot V_a\right)\cdot O\right) \tag{1.8}$$

where all of the dot-products are over feature dimensions (those of size D or D/A).

Below is pedagogical[11] sample code for such a `CausalAttention` layer[12]:

[10] However, in [8] it is instead argued that no square root should be taken in order to maximize the speed of learning via SGD.

[11] The code is written for clarity, not speed. An example optimization missing here: there is no need to form separate Q_a, K_a, V_a `Linear` layers, one large layer which is later chunked is more efficient

[12] When using sequence-parallelism, it will be more natural to separate out the final `Dropout` layer and combine it with the subsequent `LayerNorm`, as they are sharded together; see Sec. 10.4. The same is true for the `MLP` layer below.

```python
class CausalAttention(nn.Module):
    def __init__(
        self,
        block_size=K,
        dropout=0.1,
        hidden_dim=D,
        num_attn_heads=A,
    ):
        super().__init__()
        self.block_size = block_size
        self.dropout = dropout
        self.hidden_dim = hidden_dim
        self.num_attn_heads = num_attn_heads

        self.head_dim, remainder = divmod(hidden_dim, num_attn_heads)
        assert not remainder, "num_attn_heads must divide hidden_dim evenly"

        self.Q = nn.ModuleList(
            [nn.Linear(hidden_dim, self.head_dim) for _ in range(num_attn_heads)]
        )
        self.K = nn.ModuleList(
            [nn.Linear(hidden_dim, self.head_dim) for _ in range(num_attn_heads)]
        )
        self.V = nn.ModuleList(
            [nn.Linear(hidden_dim, self.head_dim) for _ in range(num_attn_heads)]
        )
        self.O = nn.Linear(hidden_dim, hidden_dim)

        self.attn_dropout = nn.Dropout(dropout)
        self.out_dropout = nn.Dropout(dropout)
        self.register_buffer(
            "causal_mask",
            torch.tril(torch.ones(block_size, block_size)[None]),
        )

    def get_qkv(self, inputs):
        queries = [q(inputs) for q in self.Q]
        keys = [k(inputs) for k in self.K]
        values = [v(inputs) for v in self.V]
        return queries, keys, values

    def get_attn_maps(self, queries, keys):
        S = queries[0].shape[1]
        norm = math.sqrt(self.head_dim)
        non_causal_attn_scores = [(q @ k.transpose(-2, -1)) / norm for q, k in zip(queries, keys)]
        # Note: this mask shape is a bit of a hack to make generation from the KV cache work without
        # specifying an extra boolean. When queries and keys have different sequence lengths and the
        # queries are of seq_len == 1,p the query attends to all of the keys; effectively there is
        # no mask at all.
        causal_attn_scores = [
            a.masked_fill(self.causal_mask[:, :S, :S] == 0, float("-inf"))
            for a in non_causal_attn_scores
        ]
        attn_maps = [a.softmax(dim=-1) for a in causal_attn_scores]
        return attn_maps
```

```
63
64      def forward(self, inputs):
65          queries, keys, values = self.get_qkv(inputs)
66          attn_maps = self.get_attn_maps(queries, keys)
67          weighted_values = torch.cat(
68              [self.attn_dropout(a) @ v for a, v in zip(attn_maps, values)], dim=-1
69          )
70          z = self.O(weighted_values)
71          z = self.out_dropout(z)
72          return z
```

The parameter count is dominated by the weight matrices which carry $4D^2$ total parameters per layer.

1.4 MLP

The feed-forward network is straightforward and corresponds to

$$z_{bsd} \leftarrow \text{Drop}\left(\phi\left(z_{bsd'}W^0_{d'e}\right)W^1_{ed}\right) \tag{1.9}$$

where W^0 and W^1 are `(D, ED)`- and `(ED, D)`-shaped matrices, respectively (see App. A for notation) and ϕ is a non-linearity[13]. In code, where we again separate out the last `Dropout` layer as we did in in Sec. 1.3.

```
6   class MLP(nn.Module):
7       def __init__(
8           self,
9           hidden_dim=D,
10          expansion_factor=E,
11          dropout=0.1,
12      ):
13          super().__init__()
14          self.hidden_dim = hidden_dim
15          self.expansion_factor = expansion_factor
16          self.dropout = dropout
17
18          linear_1 = nn.Linear(hidden_dim, expansion_factor * hidden_dim)
19          linear_2 = nn.Linear(expansion_factor * hidden_dim, hidden_dim)
20          gelu = nn.GELU()
21          self.layers = nn.Sequential(linear_1, gelu, linear_2)
22          self.dropout = nn.Dropout(dropout)
23
24      def forward(self, inputs):
25          z = self.layers(inputs)
26          z = self.dropout(z)
27          return z
```

This bock requires $2ED^2$ parameters per layer, only counting the contribution from weights.

[13] The `GeLU` non-linearity is common.

1.5 Language Model Head

The layer which converts the `(B, S, D)`-shaped outputs, z_{bsd}, to `(B, S, V)`-shaped predictions over the vocabulary, y_{bsv}, is the **Language Model Head**. It is a linear layer, whose weights are often tied to be exactly those of the initial embedding layer of Sec. 1.1.

1.6 All Together

It is then relatively straightforward to tie everything together. In code, we can first create a transformer block like

```
class TransformerBlock(nn.Module):
    def __init__(
        self,
        block_size=K,
        dropout=0.1,
        expansion_factor=E,
        hidden_dim=D,
        num_attn_heads=A,
        num_layers=L,
        vocab_size=V,
    ):
        super().__init__()
        self.block_size = block_size
        self.dropout = dropout
        self.expansion_factor = expansion_factor
        self.hidden_dim = hidden_dim
        self.num_attn_heads = num_attn_heads
        self.num_layers = num_layers
        self.vocab_size = vocab_size

        self.attn_ln = nn.LayerNorm(hidden_dim)
        self.attn = CausalAttention(
            block_size=block_size,
            dropout=dropout,
            hidden_dim=hidden_dim,
            num_attn_heads=num_attn_heads,
        )

        self.mlp_ln = nn.LayerNorm(hidden_dim)
        self.mlp = MLP(hidden_dim, expansion_factor, dropout)

    def forward(self, inputs):
        z_attn = self.attn_ln(inputs)
        z_attn = self.attn(z_attn) + inputs

        z_mlp = self.mlp_ln(z_attn)
        z_mlp = self.mlp(z_mlp) + z_attn
        return z_mlp
```

which corresponds to the schematic function

$$z \leftarrow z + \texttt{MLP}\left(\text{LayerNorm}\ (z + \texttt{CausalAttention}\left(\text{LayerNorm}\ (z)\right))\right)\ , \tag{1.10}$$

indices suppressed.

And then the entire architecture:

```
class DecoderOnly(nn.Module):
    def __init__(
        self,
        block_size=K,
        dropout=0.1,
        expansion_factor=E,
        hidden_dim=D,
        num_attn_heads=A,
        num_layers=L,
        vocab_size=V,
    ):
        super().__init__()
        self.block_size = block_size
        self.dropout = dropout
        self.expansion_factor = expansion_factor
        self.hidden_dim = hidden_dim
        self.num_attn_heads = num_attn_heads
        self.num_layers = num_layers
        self.vocab_size = vocab_size

        self.embedding = nn.Embedding(vocab_size, hidden_dim)
        self.pos_encoding = nn.Parameter(torch.randn(1, block_size, hidden_dim))
        self.drop = nn.Dropout(dropout)
        self.trans_blocks = nn.ModuleList(
            [
                TransformerBlock(
                    block_size=block_size,
                    dropout=dropout,
                    expansion_factor=expansion_factor,
                    hidden_dim=hidden_dim,
                    num_attn_heads=num_attn_heads,
                    num_layers=num_layers,
                    vocab_size=vocab_size,
                )
                for _ in range(num_layers)
            ]
        )
        self.final_ln = nn.LayerNorm(hidden_dim)
        self.lm_head = nn.Linear(hidden_dim, vocab_size, bias=False)
        self.lm_head.weight = self.embedding.weight  # Weight tying.

    def forward(self, inputs):
        S = inputs.shape[1]
        z = self.embedding(inputs) + self.pos_encoding[:, :S]
        z = self.drop(z)
        for block in self.trans_blocks:
            z = block(z)
        z = self.final_ln(z)
        z = self.lm_head(z)
        return z
```

1.7 The Loss Function

The last necessary component is the loss function. The training loop data is the `(B, K)`-shaped[14] token inputs (x_{bs}) along with their shifted-by-one relatives y_{bs} where `x[:, s + 1] == y[:, s]`. The `(B, K, V)`-shaped outputs (z_{bsv}) of the `DecoderOnly` network are treated as the logits which predict the value of the next token, given the present context:

$$p(x_{b(s+1)} = v|x_{bs}, x_{b(s-1)}, \ldots, x_{b0}) = \texttt{Softmax}_v\ z_{bsv} \tag{1.11}$$

and so the model is trained using the usual cross-entropy/maximum-likelihood loss[15]

$$\begin{aligned}\mathcal{L} &= -\frac{1}{BK}\sum_{b,s} \ln p(x_{b(s+1)} = y_{b(s+1)}|x_{bs}, x_{b(s-1)}, \ldots, x_{b0}) \\ &= \frac{-1}{BK}\sum_{b,s} \texttt{Softmax}_v\ z_{bsv}\Big|_{v=y_{b(s+1)}} \ .\end{aligned} \tag{1.12}$$

Note that the losses for all possible context lengths are included in the sum, equally weighted[16].

In `torch` code, the loss computation might look like the following (using fake data):

```
def test_loss():
    model = DecoderOnly(
        num_attn_heads=A,
        block_size=K,
        dropout=0.1,
        expansion_factor=E,
        hidden_dim=D,
        num_layers=L,
        vocab_size=V,
    )
    tokens = torch.randint(model.vocab_size, size=(B, model.block_size + 1))
    inputs, targets = tokens[:, :-1], tokens[:, 1:]
    outputs = model(inputs)
    outputs_flat, targets_flat = outputs.reshape(-1, outputs.shape[-1]), targets.reshape(-1)
    loss = F.cross_entropy(outputs_flat, targets_flat)
    assert loss
```

2 Architecture and Algorithm Variants

There are, of course, many variants on the basic architecture. Some particularly important ones are summarized here.

[14] `K` is the block size, the maximum sequence-length for the model. See App. A.

[15] Here's an alternative derivation for why this loss is minimized when the learned distribution perfectly matches the actual one. Let $p(x)$ be the actual distribution and $q_\theta(x)$ be the model. Taking the continuous case, the expected loss is $\mathcal{L} = -\int \mathrm{d}x\, p(x) \ln q_\theta(x)$. We want to minimize this, subject to the condition that $\int \mathrm{d}x\, q_\theta(x) = 1$. So, we use the calculus of variations on the loss with a Lagrange multiplier: $\mathcal{L}' = \mathcal{L} + \lambda \int \mathrm{d}x\, q_\theta(x)$. Solving $\frac{\delta \mathcal{L}'}{\delta q_\theta(x)} = 0$ yields $q_\theta(x) = p(x)$. This seems more straightforward and general than the usual argument via the KL-divergence and Jensen's inequality.

[16] In Natural Language Processing (NLP), the `perplexity` is often reported instead of the loss, which is just the exponential of the loss, a geometric-mean over the gold-answer probabilities: $\text{perplexity} = e^{\mathcal{L}} = \left(\prod_{b,s} p(x_{b(s+1)} = |x_{bs}, x_{b(s-1)}, \ldots, x_{b0})\right)^{\frac{-1}{BK}}$.

2.1 GLU Variants

In [9], Shazeer advocated for replacing the usual linear-then-activation function pattern,

$$z_{d'} = \phi\left(W_{d'd}x_d\right) \tag{2.1}$$

to

$$z_{d'} = V_{d'e}x_e\phi\left(W_{d'd}x_d\right) \ . \tag{2.2}$$

So, just perform another linear operation on the original input and broadcast it against the usual activation function output. Biases for can also be included. This construction is typically called "ϕGLU" where ϕ is the name of the activation function: ReGLU, SwiGLU/SiGLU ($\phi = x\sigma(x)$ used in the LLaMA models), etc.

2.2 Multi-Query Attention

In [7], the A different key and value matrices are replaced by a single matrix each, while A different query-heads remain. The mechanisms are otherwise unchanged: where there were previously distinct key and value tensors used across different heads, we just use the same tensors everywhere. This is **Multi-Query Attention** (MQA).

The primary reason for multi-query attention is that it vastly reduces the size of the kv-cache (see Sec. 19) during inference time, decreasing the memory-burden of the cache by a factor of A. This strategy also reduces activation memory during training, but that is more of a side-effect.

2.3 Grouped Attention

Grouped Query Attention (GQA) [10] is the natural extension of multi-query-attention to using $1 < G < A$ matrices for key and value generation. Each of the G different keys gets matched up with A/G heads (nice divisibility assumed)[17].

2.4 Parallel `MLP` and `CausalAttention` Layers

Rather than first pass inputs into the `CausalAttention` layer of each block, and then pass those outputs on to `MLP` in series, GPT-J-6B instead processes the `LayerNorm` outputs in *parallel*. That is, instead of something like

$$z \leftarrow z + \texttt{MLP}\left(\texttt{LayerNorm}\left(z + \texttt{CausalAttention}\left(z\right)\right)\right) \tag{2.3}$$

we instead have[18]

$$z \leftarrow z + \texttt{MLP}\left(z\right) + \texttt{CausalAttention}\left(z\right) \ . \tag{2.4}$$

Note that a `LayerNorm` instance is also removed.

[17] Llama-2 [11] uses GQA with $G = 8$, seemingly chosen so that each group can be sharded and put on its own GPU within a standard 8-GPU node.

[18] This alternative layer was also used in PaLM [12] where it was claimed that this formulation is $\sim 15\%$ faster due to the ability to fuse the `MLP` and `CausalAttention` matrix multiplies together (though this is not done in the GPT-J-6B repo above).

2.5 RoPE Embeddings

A shortcoming of traditional embeddings $x_{bsd} \longrightarrow x_{bsd} + p_{sd}$ is that they do not generalize very well: a model trained on such embeddings with a maximum sequence length K will do very poorly when evaluated on longer sequences. RoPE (Rotary Position Embedding) [13] and variants thereof can extend the viable context length by more clever mechanisms with stronger implicit biases.

RoPE and its variants can be motivated by a few natural conditions. Given the queries and keys for an input q_{sd}, k_{sd} (suppressing batch indices), the corresponding attention scores computation $a_{ss'}(q_s, k_{s'})$ should reasonably satisfy the below:

1. The attention score should only depend on the position indices s, s' through their difference $s - s'$, i.e., through their relative distance to each other.

2. The score computation should still be efficient, i.e., based on matrix-mulitiplies.

3. The operation should preserve the scale of the intermediate representations and attention scores, in order to avoid issues with standard normalization.

These conditions suggest a very natural family of solutions: just rotate the usual queries by some fixed element of $SO(d)$ using a generator proportional to the position index and rotate the keys by the conjugate element. That is, replace the q_{sd}, k_{sd} by

$$\begin{aligned} q'_{sd} &\equiv \left[e^{is\hat{n}\cdot T}\right]_{dd'} q_{sd'} \equiv R(s)_{dd'} q_{sd'} \\ k'_{sd} &\equiv \left[e^{-is\hat{n}\cdot T}\right]_{dd'} k_{sd'} \equiv R(s)^{\dagger}_{dd'} k_{sd'} \ , \end{aligned} \tag{2.5}$$

which makes their dot-product is $q'_{sd}k'_{s'd} = R(s - s')q_{sd}k_{sd'}$.

Performing the above computation with a dense element of $SO(D)$ is infeasible, as it would require a new dense matrix-multiply by a unique $D \times D$ matrix at each sequence position[19] In the original RoPE paper, the rotation $\hat{n}$ was chosen such that the matrices are 2×2 block-diagonal with the entries of the form[20]

$$R(s)_{[d:d+2][d:d+2]} = \begin{pmatrix} \cos(s\theta_d) & -\sin(s\theta_d) \\ \sin(s\theta_d) & \cos(s\theta_d) \end{pmatrix} \tag{2.6}$$

where

$$\theta_d = 10^{-8d/D} \ . \tag{2.7}$$

The RoPE memory costs are thus $\mathcal{O}(KD)$[21]. The sparsity present in this constrained form of the RoPE matrices means that (2.5) can be computed in $\mathcal{O}(BSD)$ time, rather than $\mathcal{O}\left(BSD^2\right)$, as it would be for a general rotation matrix. See the paper for explicit expressions.

[19] For one, the $\mathcal{O}\left(SD^2\right)$ memory cost to store the matrices would be prohibitive. The FLOPs cost is only $2BSD^2$, the same as for other matrix multiplies, but because different matrices are needed at position (it's a batched matrix multiply), these FLOPs would be much more GPU memory-bandwidth intensive.

[20] If D isn't even, the vectors are padded by an extra zero.

[21] A single RoPE buffer can be shared amongst all attention layers, amortizing the memory costs.

2.6 Flash Attention

Flash Attention [14, 15] optimizes the self attention computation by never materializing the $\mathcal{O}\left(S^2\right)$ attention scores in off-chip memory. This increases the arithmetic intensity of the computation and reduces the activation memory required, at the expense of needing recomputation in the backwards pass.

The central idea is to decompose the attention computation in the following way. Dropping the batch index, let q_{sd}, k_{sd}, v_{sd} be the queries, keys, and values, and z_{sd} be the final output. Splitting into attention heads as in $q_{sd} = q_{s(ah)} \longrightarrow q_{sah}$ and similar, the computation is[22]

$$z_{sah} = \texttt{Softmax}_{s'}\left(q_{sah'} k_{s'ah'}\right) v_{s'ah} \tag{2.8}$$

which is then concatenated as $z_{s(ah)} \longrightarrow z_{sd}$ to get the result. We are omitting the (very important) causal mask for clarity of presentation. Because each attention head computation is identical, we also omit the a-index going forward in this section.

The issue is that a naive computation would compute all $\mathcal{O}\left(S^2\right)$ components of the attention scores $q_{sh'} k_{s'h'}$ for each attention head and their exponential all at once, which incurs a penalty of shuttling back and forth $\mathcal{O}\left(S^2\right)$ elements to and from on-chip memory multiple times in order to get the final z_{sh} outputs (in addition to being potentially memory expensive). Flash Attention functions by instead computing the exponentials in stages with fewer memory transfers and never populating the attention scores or exponentials on off-chip memory.

This works by first chunking all of the inputs along their sequence dimensions as in:

- $q_{sh} = q_{(ir)h} \longrightarrow q_{irh}$ where $i \in \{0, \ldots, I-1\}$ and $r \in \{0, \ldots, R-1\}$ with $S = RI$
- $k_{sh} = k_{(jc)h} \longrightarrow k_{jch}, v_{sh} = v_{(jc)h} \longrightarrow v_{jch}$ where $j \in \{0, \ldots, J-1\}$ and $c \in \{0, \ldots, C-1\}$ with $S = JC$

The chunk sizes are determined by memory constraints, as discussed below. Then, the per-attention-head computation is equivalently written as

$$\begin{aligned}
z_{irh} &= \texttt{Softmax}_{jc}\left(q_{irh'} k_{jch'}\right) v_{jch} \\
&= \frac{\exp\left(q_{irh'} k_{jch'}\right)}{\sum_{jc} \exp\left(q_{irh''} k_{jch''}\right)} v_{jch} \\
&\equiv \frac{\sum_j Z_{irjh}}{\sum_{j'c} \exp\left(q_{irh''} k_{j'ch''}\right)} \\
&\equiv \frac{\sum_j Z_{irjh}}{\sum_{j'} L_{ij'r}} \\
&\equiv \frac{Z_{irh}}{L_{ir}}
\end{aligned} \tag{2.9}$$

[22] We omit the usual $\sqrt{D/A}$ normalization factor inside the Softmax to de-clutter the presentation. Really, this normalization should just be enforced at the level of the matrices which are used to generate the queries, keys, and values, anyway.

where we introduced the notation which will be used in the algorithm below. The algorithm proceeds similarly to how it's outlined above: we compute in chunks, looping over i and an inner j loop which is used to compute the numerator and denominator simultaneously.

Ignoring the important causal mask and not tracking the maximum logits (which we should do for numerical stability), the basic version which captures the essentials of the algorithm is below. Additional recomputation is needed for the backwards pass.

Algorithm 1 Flash Attention (Naive - Missing causal mask/max tracking.)

for $i \in \ldots$ **do** ▷ Computing outputs z_{irh} $\forall r, h$
  Initialize off-chip tensor z_{irh} to zeros
  Move q_{irh} on-chip, instantiate temp Z_{irh} to zeros on-chip.
  for $j \in \ldots$ **do** ▷ All on-chip computations. r, c indices processed in parallel.
    Move k_{jch}, v_{jch} on-chip
    $Z_{irh} \leftarrow Z_{irh} + \exp\left(q_{irh'} k_{jch'}\right) v_{jch}$ ▷ Update numerator
    $L_{ir} \leftarrow L_{ir} + \sum_c \exp\left(q_{irh'} k_{jch'}\right)$ ▷ Update denominator
  $z_{irh} \leftarrow \frac{Z_{irh}}{L_{ir}}$ ▷ Write result off-chip

We now analyze the memory transfer costs. As a baseline, vanilla attention requires $\mathcal{O}\left(S^2 + DS\right)$ memory transfers per attention head, where the two factors come from the attention scores and q, k, v, respectively. For flash attention, we no longer shuttle the attention scores off-chip, but k, v are repeatedly moved back and forth. These transfers form most of the memory operations in the inner loop above, which access $\mathcal{O}\left(IJCH\right) \sim \mathcal{O}\left(\frac{HS^2}{R}\right)$ elements over the lifetime of the algorithm (per attention head). The factor H/R determines the memory-access advantage, and this number is bound by the on-chip memory size. The on-chip bytes from the queries, keys, and vectors take $\mathcal{O}\left(CH + RH\right)$ memory and the temporaries from attention scores and exponentials require $\mathcal{O}\left(RC\right)$. If we have M bytes of on-chip memory, then we have the constraint $CH + RH + RC \lesssim M$, and assuming assuming the chunks were chosen to maximize on-chip memory usage, $\frac{H}{R} \sim \frac{H^2}{M}$. Since $M \sim 10^5$ bytes on 2023 GPUs, this is a small factor for the typical head dimensions $H \sim 64$, as desired.

Flash attention is also a big win for activation memory: a naive algorithm has a $\mathcal{O}\left(ABS^2\right)$ per-layer contribution to activation memory due to needing to save the attention weights, but these are discarded and re-computed for flash attention. The only additional memory cost comes from the $\mathcal{O}\left(ABS\right)$ elements in the ℓ_{abs} statistics, which are dominated by the $\mathcal{O}\left(BSD\right)$ costs from needing to save inputs, and hence negligible.

2.6.1 The Details

Here we give more detailed descriptions of the flash-attention forwards and backwards passes.

For the forwards pass, we add in maximum-logits tracking for more numerically stable exponential computation and the causal mask. The causal mask $C_{ss'} = C_{(ir)(jc)}$ is zero if $s \geq s'$ and $-\infty$ otherwise. The algorithm is as below.

For the backwards pass, the main complication comes from computing derivatives with respect to the attention scores. Recalling the `Softmax` derivative (6.10), given gradients $\frac{\partial \mathcal{L}}{\partial z_{irj}} \equiv g_{irj}$ we

Algorithm 2 Flash Attention Forward Pass

for $i \in \ldots$ **do**
  Initialize off-chip tensors z_{irh}, ℓ_{ir} to zeros ▷ Computing outputs $z_{irh}\ \forall r, h$
  Move q_{irh} on-chip, instantiate temp Z_{irh} to zeros and $M_{ir}^{\text{new}}, M_{ir}^{\text{old}}$ to $-\infty$ on-chip
  for $j \in \ldots$ **do** ▷ All on-chip computations. r, c indices processed in parallel.
    Move k_{jch}, v_{jch} on-chip
    $S_{irjc} \leftarrow q_{irh'} k_{jch'} + C_{ijrc}$ ▷ `Softmax` logits + causal mask
    $M_{ir}^{\text{new}} \leftarrow \max\left(M_{ir}^{\text{old}}, \max_c S_{irjc}\right)$
    $Z_{irh} \leftarrow Z_{irh} + \exp\left(S_{ijrc} - M_{ir}^{\text{new}}\right) v_{jch}$ ▷ Update numerator
    $L_{ir} \leftarrow e^{M_{ir}^{\text{old}} - M_{ir}^{\text{new}}} L_{ir} + \sum_c \exp\left(S_{ijrc} - M_{ir}^{\text{new}}\right)$ ▷ Update denominator
    $M_{ir}^{\text{old}} \leftarrow M_{ir}^{\text{new}}$
  $z_{irh} \leftarrow \frac{Z_{irh}}{L_{ir}}$, $\ell_{ir} \leftarrow M_{ir}^{\text{old}} + \ln L_{ir}$ ▷ Write results off-chip. ℓ_{ir} for backwards

have the building blocks[23]

$$\begin{aligned}
\frac{\partial P_{irjc}}{\partial S_{irj'c'}} &= P_{irjc}\delta_{jj'}\delta_{cc'} - P_{ijrc}P_{irj'c'} \\
\frac{\partial \mathcal{L}}{\partial P_{irjc}} &= g_{irh} v_{jch} \\
\frac{\partial \mathcal{L}}{\partial S_{irjc}} &= g_{irh} \frac{\partial P_{irjc}}{\partial S_{irj'c'}} \\
&= g_{irh}\left(P_{irjc} v_{jch} - P_{irjc} P_{irj'c'} v_{j'c'h}\right) \\
&= g_{irh}\left(P_{irjc} v_{jch} - P_{irjc} z_{irh}\right) \\
&= P_{irjc}\left(\frac{\partial \mathcal{L}}{\partial P_{irjc}} - g_{irh} z_{irh}\right)
\end{aligned} \tag{2.10}$$

from which we compute

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial v_{jch}} &= g_{irh} P_{irjc} \\
\frac{\partial \mathcal{L}}{\partial q_{irh}} &= \frac{\partial \mathcal{L}}{\partial S_{irjc}} k_{jch} \\
\frac{\partial \mathcal{L}}{\partial k_{jch}} &= \frac{\partial \mathcal{L}}{\partial S_{irjc}} q_{irh}
\end{aligned} \tag{2.11}$$

Above we let $P_{ijrc} \equiv \texttt{Softmax}_{jc}(S_{irjc})$ where $S_{ijrc} \equiv q_{irh'} k_{jch'} + C_{ijrc}$, keeping notation similar to the above algorithm. All of this suggests a very similar algorithm to the above. Using the unfortunate, but common, notation, $\mathrm{d}X = \frac{\partial \mathcal{L}}{\partial X}$ the algorithm is[24]:

[23] The fact that we can replace the j' sum with the cached attention outputs in the final derivative below is crucial.

[24] In the FA2 paper, they actually pre-compute the $g_{irh} z_{irh}$ sum prior to the main loop and store it in a tensor they call D. And in the official `triton` example, $\mathrm{d}q$ is computed in a separate loop. So, take the below as more of a guideline than a strict recipe.

Algorithm 3 Flash Attention Backward Pass

for $i \in \ldots$ **do**
  Initialize off-chip tensors $\mathrm{d}q_{irh}, \mathrm{d}k_{jch}, \mathrm{d}v_{jch}$ to zeros
  Move $z_{irh}, q_{irh}, g_{irh}$ and the cached ℓ_{ir} on-chip.
  for $j \in \ldots$ **do** ▷ All on-chip computations. r, c indices processed in parallel.
    Instantiate $P_{irjc}, \mathrm{d}P_{irjc}, \mathrm{d}S_{irjc}$ to zeros.
    Move k_{jch}, v_{jch} on-chip
    $P_{irjc} \leftarrow \exp\left(q_{irh'}k_{jch'} + C_{ijrc} + \ell_{ir}\right)$ ▷ Get probabilities.
    $\mathrm{d}P_{irjc} \leftarrow g_{irh}v_{jch}$ ▷ Get derivatives w.r.t. P
    $\mathrm{d}S_{irjc} \leftarrow P_{ijrc}\left(\mathrm{d}P_{ijrc} - g_{irh}z_{irh}\right)$ ▷ Get derivatives w.r.t. S
    $\mathrm{d}k_{jch} \leftarrow \mathrm{d}S_{irjc}q_{irh}$ ▷ Derivatives w.r.t. $\{q, k, v\}$
    $\mathrm{d}v_{jch} \leftarrow g_{irh}P_{irjc}$
    Write $\mathrm{d}k, \mathrm{d}v$ derivatives to off-chip
    $\mathrm{d}q_{irh} \leftarrow \mathrm{d}q_{irh} + \mathrm{d}S_{irjc}k_{jch}$
  Write $\mathrm{d}q$ derivative to off-chip

2.7 Linear Attention

Linear attention [16] removes the `Softmax` operation in the attention layer in order to reduce the inference costs in terms of compute and time, both.

To review, the (single-head) attention operation is

$$\begin{aligned} z_{sd} &= \texttt{Softmax}_{s'}\left(q_{sd'}k_{s'd'}\right)v_{s'd} \\ &\equiv A_{ss'}v_{s'd} \ . \end{aligned} \tag{2.12}$$

In order to generate the final, $s = -1$ token using a kv-cache, generation time then requires reading in $\mathcal{O}\left(DS\right)$ bytes and performing $\mathcal{O}\left(DS\right)$ operations to generate the new token. However, if we remove the `Softmax`, then we can write the above as

$$\begin{aligned} z_{sd} &= q_{sd'}k_{s'd'}v_{s'd} \\ &= q_{sd'}B_{d'd} \ . \end{aligned} \tag{2.13}$$

This would let us cache the $\mathcal{O}\left(D^2\right)$ $B_{d'd}$ matrix and generating the next token only takes $\mathcal{O}\left(D^2\right)$ operations, an $\mathcal{O}\left(D/S\right)$ improvement on both fronts.

The essential point is that for standard attention, the entire $A_{ss'}$ matrix must be computed anew for each new token, while $B_{dd'}$ can instead be iteratively updated via a cheap computation.

The causal masking looks a little different for linear attention. The causal mask is not needed during vanilla next-token generation, but is for parallelized training. Computing of the z_{sd} in parallel, as in training, requires generating S different matrices $B^s_{d'd}$, one for each token position: $B^s_{d'd} = \texttt{cumsum}_s\left(k_{sd'}v_{sd}\right)$, effectively. Flash-attention-like techniques can be used to avoid materializing all of the $\mathcal{O}\left(SD^2\right)$ elements at once.

Part II
State Space Models

3 Intro

Needing to re-reference the entire previously-generated prefix at generation time is a major pain point for transformers models. Token generation is $\mathcal{O}(S)$ State space models return, more or less, to the old LSTM type strategy of encoding the conditional history which informs generation into a finite-sized state. The dream is faster generation and better memory efficiency.

4 S4

The S4 model of [17] is a good starting point. These are based off a continuous representation in which some input signal[25] $x_a(t)$ is converted to an output $y_c(t)$ via an in intermediate latent variable $h_b(t)$, with the above related as in

$$\begin{aligned} \partial_t h_b(t) &= A_{bb'} h_{b'}(t) + B_{ba} x_a(t) \\ y_c(t) &= C_{hb} h_b(t) + D_{ca} x_a(t) \ . \end{aligned} \tag{4.1}$$

The capitalized tensors are the learnable weight matrices. D is often set to zero in the literature. Basically, the information in the sequence x_s is stored in h_s, an internal memory for the model, much like the RNN/LSTM models of the past.

For discrete sequences, we discretize:

$$\begin{aligned} h_{bs} &= A_{bb'} h_{b'(s-1)} + B_{ba} x_{as} \\ y_{cs} &= C_{cb} h_{bs} + D_{ca} x_{as} \ . \end{aligned} \tag{4.2}$$

where one can also relate these weights to those in (4.1) given the discretization scheme (see 5).

Subject to the initial condition $h_b^{-1} = 0$, the above solves to

$$y_s = \sum_{s'=0}^{s} C \cdot A^{s-s} \cdot B \cdot x_{s'} + D x_s \ , \tag{4.3}$$

omitting hidden dimension indices. Proper normalization of the various weights is non-trivial; see [17] for details. Further, diagonalization clearly makes the A^{s-n} computation easier, but care must be taken here, too. Clearly, the above computation is highly parallelizable. The S4 (and mamba) papers describe (4.3) as a

Writing the above operation as $y_{cs} = \Sigma_{cass'} x_{as'}$, one can build an non-linear S4 layer by acting on the output with a non-linearity and then mixing feature dimensions with a weight matrix:

$$z_{cs} = W_{cc'} \phi\left(\Sigma^{c'ass'} x_{as'}\right) \tag{4.4}$$

[25] We use the notation of the mamba paper [18], which differs from that of the S4 paper [17].

Assuming the c and a hidden dimensions have the same size, the operations can then be naturally composed.

Taking all hidden dimensions to have size $\mathcal{O}(D)$, the number of learnable weights is $\mathcal{O}\left(D^2\right)$. Training can be parallelized across the sequence dimension (via the representation (4.3), scaling linearly in sequence length. Iterative generation from $x_{as} \longrightarrow y_{cs}$, given knowledge of the previous hidden state $h_{b(s-1)}$ takes only $\mathcal{O}\left(D^2\right)$ (via the representation (4.2)). There is no sequence-length dependence for next-output generation, unlike for transformers, which is the main draw here: constant-time generation.

5 Mamba

A large limitation of the S4 model (4.2) is that the various weights are fixed quantities which do not adjust to the input[26] x_{sd}. Mamba [18] extends S4 by replacing the fixed weights by functions of the inputs. This destroys the recursive structure and requires various techniques for an efficient GPU implementation, which is the primary focus of the paper.

The mamba architecture is as follows, based on the implementation in `mamba.py` and `mamba_ssm`. Notation for dimensions and tensors:

- Mamba maps sequences to sequences, the same as for transformers. $z_{sd} = \texttt{mamba}\,(x_{sd})$. Batch dimension suppressed throughout.
- Various dimensions:
  - $d \in \{0, \ldots, D-1\}$: the input's hidden dimensions, `d_model`.
  - $e \in \{0, \ldots, E \times D - 1\}$: expanded internal hidden dimension. Usually $E = 2$ in practice.
  - $s \in \{0, \ldots, S-1\}$: sequence length.
  - $n \in \{0, \ldots, N-1\}$: another internal hidden dimension, controlling the size of the internal memory; `d_state`. Defaults to 16.
  - $r \in \{0, \ldots, R-1\}$: another internal hidden dimension, `d_rank`. Defaults to $\lceil D/16 \rceil$.
  - $c \in \{0, \ldots, C-1\}$: convolution kernel size; `d_conv`, 4 by default. Used to convolve over the sequence dimension.
- Learnable parameters[27]:
  - Two in-projectors from `d_model` to the expanded dimension: $W^{I_0}_{ed}$, $W^{I_1}_{ed}$.
  - Out-projector from the expanded internal dimension back to `d_model` W^{O}_{de}.

[26] For instance, we could ask our architecture to process two independent sequences concatenated together with a special separator token in the middle. The hidden state should be reset at the separator token and the mamba architecture would be (in-principle) capable of this, while the S4 would not.

[27] In practice, many of these are fused together for more efficient matmuls. We also omit potential bias terms.

– Two projectors used in creating the intermediate Δ_{se}: $W^{\Delta_0}_{re}$, $W^{\Delta_1}_{er}$.

– Projectors for creating the intermediates B_{sn} and C_{sn}: W^B_{ne}, W^C_{ne}

– Convolutional kernel W^K_{ec}.

– Selective-scan weights W^A_{en}.

– Residual connection weights W^D_e.

The notation here is not the same as that of the papers. We write all learnable weights as $W^X_{\dots}$.

Mamba blocks then perform the following logical operation:

Algorithm 4 Mamba

Inputs: tensor $x_{sd} \in \mathbb{R}^{S\times D}$
$x^0_{se} = W^{I_0}_{ed} x_{sd}$, $x^1_{se} = W^{I_1}_{ed} x_{sd}$ ▷ Create expanded tensors from inputs (can fuse)
$x^2_{se} = K_{ess'} \star x^1_{se}$ ▷ 1D grouped convolution over the sequence dimension using W^K_{ec}.
$x^3_{se} = \phi\left(x^2_{se}\right)$ ▷ Elementwise non-linearity (`F.silu` default)
$x^4_{se} = \texttt{selective\_scan}\left(x^3_{se}\right)$ ▷ Selective scan (see below).
$x^5_{se} = x^4_{se} \otimes \phi\left(x^0_{se}\right)$ ▷ Elementwise product and non-linearity (`F.silu` default)
$z_{sd} = W^O_{de} x^5_{se}$ ▷ Project back down.
return $z_{sd} \in \mathbb{R}^{S\times D}$

where `selective_scan` operation is the above is[28]

Algorithm 5 Selective Scan: `selective_scan`

Inputs: tensor $x_{se} \in \mathbb{R}^{S\times E}$
$B_{sn} = W^B_{ne} x_{se}$ ▷ Create intermediates B, C, Δ (can fuse).
$C_{sn} = W^C_e x_{se}$
$\Delta_{se} = W^{\Delta_1}_{er} W^{\Delta_0}_{re} x_{se}$.
Solve recursion, subject to $h_{(-1)en} = 0$:

$$\begin{aligned} h_{sen} &= \exp\left(\Delta_{se} W^A_{en}\right) h_{(s-1)en} + \Delta_{se} B_{sn} x_{se} \\ y_{se} &= C_{sn} h_{sen} + W^D_e x_{se} \\ \implies y_{se} &= C_{sn}\left(\sum_{s'=0}^{s} e^{\Delta_{se} W^A_{en}} \times \ldots \times e^{\Delta_{s'e} W^A_{en}} \Delta_{s'e} B_{s'n} x_{s'e}\right) + W^D_e x_{se} \end{aligned} \tag{5.1}$$

return $y_{se} \in \mathbb{R}^{S\times E}$

As noted above, the creation of the intermediates $x^0_{se}, x^1_{se}, B_{sn}, C_{sn}$ and part of Δ_{se} can all be formed in a single large matmul.

[28] The `mamba_ssm` and `mamba.py` implementations differ in the first step in that the latter optionally applies a norm operator post-projection. The exponentials here might seem odd, but are probably motivated by the existence of good cumulative sum kernels, which is how the exponents can be computed.

5.1 Mamba 2

Mamba2 introduces some changes:

- The n-dimension is expanded to `ngroups` such dimensions (though `ngroups`=1 is the default), with associated index $g \in \{0, \ldots, G-1\}$, $G \equiv$ `ngroups`. Adding a non-trivial `ngroups` seems completely degenerate with expanding the n dimension of size `d_state` to size `d_state` $\times$ `ngroups`.
- A head-index $a \in \{0, \ldots, A-1\}$ ($A \equiv$ `nheads`) and head dimension $h \in \{0, \ldots, H\}$ ($A \times H = E$) are introduced, analogously to transformers.
- The e-index from two selective-scan weights is removed: they are now per-head scalars W^A_a, W^D_a.
- The intermediate Δ_{sa} is also reduced to a per-head, per-sequence-position scalar, with respect to the hidden dimension. This tensor is now created via a single matmul with weight W^{Δ}_{ae}.
- The short 1D convolution is now also taken over the B and C intermediates with kernels $W^{K_B}_{gnc}$, $W^{K_B}_{gnc}$.

The updated model:

Algorithm 6 Mamba2

Inputs: tensor $x_{sd} \in \mathbb{R}^{S \times D}$
$x^0_{se} = W^{I_0}_{ed} x_{sd}$, $x^1_{se} = W^{I_1}_{ed} x_{sd}$ ▷ Create expanded tensors from inputs (can fuse)
$x^2_{se} = K_{ess'} \star x^1_{se}$ ▷ 1D grouped convolution over the sequence dimension (fused)
$x^3_{se} = \phi\left(x^2_{se}\right)$ ▷ Elementwise non-linearity (`F.silu` default)
$x^4_{se} =$ `selective_scan2` $\left(x^3_{se}\right)$ ▷ Selective scan (see below).
$x^5_{se} = \mathrm{Norm}\left(x^4\right)_{se} \otimes \phi\left(x^0_{se}\right)$ ▷ Elementwise product, non-linearity, and norm (`RMS` default)
$z_{sd} = W^O_{de} x^5_{se}$ ▷ Project back down.
return $z_{sd} \in \mathbb{R}^{S \times D}$

The mechanical differences are the normalization step and the details of the `selective_scan2` operation, which is essentially the same as before, but now the hidden e is split into multiple attention heads, analogously to transformer models:

As before, many of the matmuls can be performed as one big operation, and the three short convolutions can be similarly fused into a single convolution. The two algorithms Algo. 6 and Algo. 11 are nearly identical; they just differ in some tensor shapes.

5.2 Mamba2 Duality with Attention

There are only two steps in which tokens at different temporal positions interact in the Mamba2 model:

1. In the short 1D convolution.

Algorithm 7 Selective Scan 2: `selective_scan2`

Inputs: tensor $x_{se} \in \mathbb{R}^{S\times E}$

$x_{sah} = x_{s(ah)}$ ▷ Break the inputs up into attention heads.

$B_{sgn} = W^B_{gne} x_{se}$ ▷ Create intermediates B, C, Δ (can fuse)[29].

$C_{sgn} = W^C_{gne} x_{se}$

$\Delta_{sa} = W^{\Delta}_{ae} x_{se}$.

$\Delta_{sa} = \texttt{Softplus}\left(\Delta_{sa}\right)$. ▷ For some reason. $\texttt{Softplus}(x) \equiv \ln\left(1 + e^x\right)$.

$B_{sgn} = K^B_{gnss'} \star B_{sgn}$ ▷ 1D grouped convolution over the sequence dimension (fused)

$C_{sgn} = K^C_{gnss'} \star C_{sgn}$

Solve recursion, subject to $h_{(-1)gahn} = 0$:

$$\begin{aligned} h_{sgahn} &= \exp\left(\Delta_{sa} W^A_a\right) h_{(s-1)gahn} + \Delta_{sa} B_{sgn} x_{se} \\ y_{sah} &= C_{sgn} h_{sgahn} + W^D_a x_{sah} \\ \implies y_{sah} &= C_{sgn} \left(\sum_{s'=0}^{s} e^{\Delta_{sa} W^A_a} \times \ldots \times e^{\Delta_{s'a} W^A_a} \Delta_{s'a} B_{s'gn} x_{s'ah} \right) + W^D_a x_{sah} \end{aligned} \tag{5.2}$$

$y_{se} = y_{s(ah)}$ ▷ Concatenate the heads back together.

return $y_{se} \in \mathbb{R}^{S\times E}$

2. In the recurrence relation, where we create the intermediate

$$z_{sah} = C_{sgn} \left(\sum_{s'=0}^{s} e^{\Delta_{sa} W^A_a} \times \ldots \times e^{\Delta_{s'a} W^A_a} \Delta_{s'a} B_{s'gn} x_{s'ah} \right) \equiv M_{ass'} x_{s'ah} \tag{5.3}$$

which is the most complicated step of the model.

As noted above, the second case is ultimately just a matrix-multiply on the input tensors x_{sah} with the tensor $M_{ass'}$, where operations across attention head are all independent. The $M_{ass'}$ tensor has $\mathcal{O}\left(AS^2\right)$ elements, which we clearly do not want to concurrently materalize. All of this should sound familiar: the above is exactly analogous to the structure and problems of flash attention, Sec. 2.6, the only difference begin how the linear operator $M_{ass'}$ is constructed: $M_{ass'} = \texttt{Softmax}\left(q^a_{sh} k^a_{s'h}\right)$ in standard attention, and as above for Mamba2, say. This is the "duality" discussed in [19] and the broad strokes of the efficient algorithm implementation for Mamba2 echos that of flash attention: partition the computation over the sequence dimensions and compute z_{sah} in chunks over the s-dimension, so as to avoid realizing any $\mathcal{O}\left(S^2\right)$ tensors.

Similar statements hold for the original Mamba; the index names and choices just make the analogy more readily recognizable in Mamba2.

5.3 Aren't These Just RNNs?

Yes, but very special ones with the important computational difference that the recursion relations are *linear* in the hidden state h. This crucial difference makes it possible to parallelize the operations during training. Compare (4.2) to what typical RNN recursion relations would look like:

$$h_{bs} = \phi\left(A_{bb'} h_{b'(s-1)} + B_{ba} x_{as}\right)$$

$$y_{cs} = \phi\left(C_{cb} h_{bs} + D_{ca} x_{as}\right) \ . \tag{5.4}$$

for some non-linearity ϕ. The recursion relations would solve to an expression with nested ϕ factors which would make the computation of h_{bs} non-associative. But in the linear $\phi(x) = x$ limit, the operations are *associative* which makes them *parallelizable*, via known scan algorithms [20].

Part III
Training

6 Memory

In this section we summarize the train-time memory costs of Transformers under various training strategies[30].

The memory cost is much more than simply the cost of the model parameters. Significant factors include:

- Optimizer states, like those of `Adam`
- Mixed precision training costs, due to keeping multiple model copies.
- Gradients
- Activation memory[31] , needed for backpropagation.

Because the activation counting is a little more involved, it is in its own section.

Essentials

Memory costs count the elements of all tensors in some fashion, both from model parameters and intermediate representations. The gradient and optimizer state costs scale with the former quantity: $\mathcal{O}(N_{\text{params}}) \sim \mathcal{O}(LD^2)$, only counting the dominant contributions from weight matrices. Activation memory scales with the latter, which for a `(B, S, D)`-shaped input gives $\mathcal{O}(BDLS)$ contributions from tensors which preserve the input shape, as well as $\mathcal{O}(ABLS^2)$ factors from attention matrices.

6.1 No Sharding

Start with the simplest case where there is no sharding of the model states. Handling the different parallelism strategies later will be relatively straightforward, as it involves inserting just a few factors here and there.

6.1.1 Parameters, Gradients, Optimizer States, and Mixed Precision

Memory from the bare parameter cost, gradients, and optimizer states are fixed costs independent of batch size and sequence-length (unlike activation memory), so we discuss them all together here. The parameter and optimizer costs are also sensitive to whether or not mixed-precision is used, hence we also address that topic, briefly. We will assume the use of `Adam`[32] throughout, for simplicity

[30] A nice related blog post is here.

[31] Activations refers to any intermediate value which needs to be cached in order to compute backpropagation. We will be conservative and assume that the inputs of all operations need to be stored, though in practice gradient checkpointing and recomputation allow one to trade caching for redundant compute. In particular, flash attention [14] makes use of this strategy.

[32] Which stores two different running averages per-model parameter.

and concreteness. It will sometimes be useful below to let p to denote the precision in bytes that any given element is stored in, so `torch.float32` corresponds to $p = 4$, for instance. Ultimately, we primarily consider vanilla training in $p = 4$ precision and `torch.float32`/`torch.float16` ($p = 4/p = 2$) mixed-precision, other, increasingly popular variants to exist, so we keep the precision variable where we can.

Without mixed precision, the total cost of the `torch.float32` ($p = 4$ bytes) model and optimizer states in bytes is then

$$M_{\text{model}} = 4N_{\text{params}} \ , \quad M_{\text{optim}} = 8N_{\text{params}} \quad \text{(no mixed precision,} p = 4\text{)} \tag{6.1}$$

where, from the previous section, the pure parameter-count of the decoder-only Transformers architecture is

$$N_{\text{params}} \approx (4 + 2E)LD^2 \times \left(1 + \mathcal{O}\left(\frac{V}{DL}\right) + \mathcal{O}\left(\frac{1}{D}\right)\right) \ . \tag{6.2}$$

where the first term comes from the `TransformerBlock` weight matrices[33], the first omitted subleading correction term is the embedding matrix, and the last comes from biases, `LayerNorm` instances, and other negligible factors. The optimizer states cost double the model itself.

The situation is more complicated when mixed-precision is used [21]. The pertinent components of mixed-precision[34]:

- A half-precision ($p = 2$ bytes) copy of the model is used to perform the forwards and backwards passes
- A second, "master copy" of the model is also kept with weights in full $p = 4$ precision
- The internal `Adam` states are kept in full-precision

Confusingly, the master copy weights are usually accounted for as part of the optimizer state, in which case the above is altered to

$$M_{\text{model}} = 2N_{\text{params}} \ , \quad M_{\text{optim}} = 12N_{\text{params}} \quad \text{(mixed precision)}. \tag{6.3}$$

The optimizer state is now six times the cost of the actual model used to process data and the costs of (6.3) are more than those of (6.1). However, as we will see, the reduced cost of activation memory can offset these increased costs, and we get the added benefit of increased speed due to specialized hardware. The above also demonstrates why training is so much more expensive than inference.

[33] So, in the usual $E = 4$ case, the `MLP` layers are twice as costly as the `CausalAttention` layers.

[34] A note on the implementation of mixed-precision in `torch`: usually mixed-precision occurs by wrapping the forward pass in a context manager, `torch.autocast`. The default behavior is to then create copies of some tensors in lower-precision and do the forward pass with those. For instance, this is done with matrix-multiplies whose arguments and outputs will be in `torch.float16`, but for sums the inputs and outputs will all be `torch.float32`, for vanilla mixed-precision usage. Consequently, any such `torch.float16` versions of tensor will often persist effectively as contributors to activation memory, since the backwards pass will need those same tensors. This can be verified by inspecting the saved tensors: if `z` is the output of a matrix-multiply in such an autocast context, `z.grad_fn._saved_mat2` will be a `torch.float16` copy of the weights used to perform the matrix-multiply. In effect, the cost of the model weights which are used for the actual forward pass are only materialized within the lifetime of the context manager.

6.1.2 Gradients

Gradients are pretty simple and always cost the same regardless of whether or not mixed-precision is used:

$$M_{\text{grad}} = 4N_{\text{params}} \ . \tag{6.4}$$

In mixed precision, even though the gradients are initially computed in $p = 2$, they have to be converted to $p = 4$ to be applied to the master weights of the same precision.

6.1.3 Activations

Activations will require a more extended analysis [5]. Unlike the above results, the activation memory will depend on both the batch size and input sequence length, B and S, scaling linearly with both.

Attention Activations We will count the number of input elements which need to be cached. Our `(B, S, D)`-shaped inputs to the attention layer with BDS elements are first converted to $3BDS$ total query, key, value elements, and the query-key dot products produce ABS^2 more, which are softmaxed into ABS^2 normalized scores. The re-weighted inputs to the final linear layer also have BDS elements, bringing the running sum to $BS\,(5D + 2AS)$

Finally, there are also the dropout layers applied to the normalized attention scores and the final output whose masks must be cached in order to backpropagate. In torch, the mask is a `torch.bool` tensor, but surprisingly these use one *byte* of memory per element, rather than one bit [35]. Given this, the total memory cost from activations is

$$M_{\text{act}}^{\texttt{Attention}} = BLS\,((5p + 1)D + (2p + 1)AS) \ . \tag{6.5}$$

MLP Activations First we pass the `(B, S, D)`-shaped inputs into the first MLP layer. These turn into the `(B, S, E*D)` inputs of the non-linearity, whose same-shaped outputs are then passed into the last `Linear` layer, summing to $(2E+1)BDS$ total elements thus far. Adding in the dropout mask, the total memory requirement across all `MLP` layers is:

$$M_{\text{act}}^{\texttt{MLP}} = (2Ep + p + 1)BDLS \ . \tag{6.6}$$

LayerNorm, Residual Connections, and Other Contributions Then the last remaining components. The `LayerNorm` instances each have BDS inputs and there are two per transformer block, so $M_{\text{act}}^{\texttt{LayerNorm}} = 2pBDLS$, and there is an additional instance at the end of the architecture[36]. There are two residual connections per block, but their inputs do not require caching (since their derivatives are independent of inputs). Then, there are additional contributions from pushing the last layer's outputs through the language-model head and computing the loss function, but these do not scale with L and are ultimately $\sim \mathcal{O}\left(\frac{V}{DL}\right)$ suppressed, so we neglect them.

[35] As you can verify via `4 * torch.tensor([True]).element_size() == torch.tensor([1.]).element_size()`.
[36] Following [5] we will neglect this in the below sum, an $\mathcal{O}\,(1/L)$ error

Total Activation Memory Summing up the contributions above, the total activation memory cost per-layer is

$$M_{\text{act}}^{\text{total}} \approx 2BDLS\left(p(E+4)+1+\mathcal{O}\left(\frac{V}{DL}\right)\right)+ABLS^2\left(2p+1\right) \ . \tag{6.7}$$

Evaluating in common limits, we have:

$$\begin{aligned} M_{\text{act}}^{\text{total}}\Big|_{E=4,p=4} &= BLS\left(66D+15AS\right) \\ M_{\text{act}}^{\text{total}}\Big|_{E=4,p=2} &= BLS\left(34D+5AS\right) \end{aligned} \tag{6.8}$$

When does mixed-precision reduce memory? (Answer: usually.) We saw in Sec. 6.1.1 that mixed precision *increases* the fixed costs of non-activation memory, but from the above we also see that it also *reduces* the activation memory and the saving increase with larger batch sizes and sequence lengths. It is straightforward to find where the tipping point is. Specializing to the case $E = 4$, vanilla mixed-precision case with no parallelism[37], the minimum batch size which leads to memory savings is

$$B_{\text{min}} = \frac{6D^2}{8DS+AS^2} \ . \tag{6.9}$$

Plugging in numbers for the typical $\mathcal{O}\left(40 \text{ GiB}\right)$ model in the Summer of 2023 gives $B_{\text{min}} \sim \mathcal{O}\left(1\right)$, so mixed-precision is indeed an overall savings at such typical scales.

[37] With both tensor- and sequence-parallelism, the parallelism degree T actually drops out in the comparison (since both form of memory are decrease by $1/T$, so this restriction can be lifted.

Side Note: Optimizations

The above analysis is conservative and accounts for more tensors than are actually saved in practice.

For instance, both the input and outputs of all non-linearities were counted, but there are many activations whose derivatives can be reconstructed from its outputs alone: $\phi'(z) = F(\phi(z))$ for some F. Examples:

- `ReLU`: since $\phi(z) = z\theta(z)$, then (defining the derivative at zero to be zero) $\phi'(z) = \theta(z) = \theta(\phi(z))$. Correspondingly, torch only uses the `ReLU` outputs to compute the derivative (there is no self arg in the `threshold_backward(grad, result, 0)` line).
- `tanh`: since $\tanh'(z) = 1 - \tanh(z)^2$.

Other cases do not have this nice property, in which case both the inputs and outputs need to be stored:

- `GeLU` [22]: $\phi(z) = z\Phi(z)$ here and the derivative $\phi'(z) = \Phi(z) + \frac{ze^{-z^2/2}}{\sqrt{2\pi}}$, both the inputs and outputs must be used in the backwards pass..

  The explicit CUDA kernel is here.

If the inputs in each of these cases are not needed for any other part of the backwards pass, they are garbage collected in `torch` soon after creation.

Example : `Softmax` is another instance where this occurs, since

$$\partial_i \texttt{Softmax}\,(x_j) = \delta_{ij}\,\texttt{Softmax}\,(x_j) - \texttt{Softmax}\,(x_i)\,\texttt{Softmax}\,(x_j) \tag{6.10}$$

Because of this, the actual amount of activation memory due to the attention layer after the forwards pass is (6.5) with $2p \longrightarrow p$ in the $\mathcal{O}\left(S^2\right)$ term, though the above expression better reflects the necessary peak memory.

6.2 Case Study: Mixed-Precision GPT3

Let's run through the numbers for mixed-precision GPT3 with parameters:

$$L = 96\ , \quad D = 12288\ , \quad A = 96\ , \quad V = 50257\ . \tag{6.11}$$

We are leaving the sequence-length unspecified, but the block-size (maximum sequence-length) is $K = 2048$.

Start by assuming no parallelism at all. The total (not per-layer!) non-activation memory is

$$M_{\text{non-act}}^{\texttt{GPT-3}} \approx 1463 \text{ TiB} \tag{6.12}$$

which can be broken down further as

$$M_{\text{params}}^{\texttt{GPT-3}} \approx 162 \text{ TiB}\ , \quad M_{\text{grads}}^{\texttt{GPT-3}} \approx 325 \text{ TiB}\ , \quad M_{\text{optim}}^{\texttt{GPT-3}} \approx 975 \text{ TiB}\ . \tag{6.13}$$

The embedding matrix only makes up $\approx .3\%$ of the total number of parameters, justifying our neglect of its contribution in preceding expressions.

The activation memory is

$$M_{\text{act}}^{\texttt{GPT-3}} \approx 3 \times 10^{-2} BS \times \left(1 + \frac{S}{10^3}\right) \text{ TiB} \ . \tag{6.14}$$

Note that the attention matrices, which are responsible for $\mathcal{O}\left(S^2\right)$ term, will provide the dominant contribution to activation memory in the usual $S \gtrsim 10^3$ regime.

In the limit where we process the max block size ($S = K = 2048$), the ratio of activation to non-activation memory is

$$\frac{M_{\text{act}}^{\texttt{GPT-3}}}{M_{\text{non-act}}^{\texttt{GPT-3}}}\Big|_{S=2048} \approx .2B \ . \tag{6.15}$$

So, the activation memory is very significant for such models.

Using tensor parallelism into the above with the maximal $T = 8$ which can be practically used, the savings are significant. The total memory is now

$$M_{\text{total}}^{\texttt{GPT-3}} \approx 187 \text{ TiB} + 10^{-2} BS + 5 \times 10^{-6} BS^2 \ . \tag{6.16}$$

7 Training FLOPs

The total number of floating point operations (FLOPs)[38] needed to process a given batch of data is effectively determined by the number of matrix multiplies needed.

Recall that a dot-product of the form $v \cdot M$ with $v \in \mathbb{R}^m$ and $M \in \mathbb{R}^{m,n}$ requires $(2m-1) \times n \approx 2mn$ FLOPs . For large language models, $m, n \sim \mathcal{O}\left(10^3\right)$, meaning that even expensive element-wise operations like `GeLU` acting on the same vector v pale in comparison by FLOPs count [39]. It is then a straightforward exercise in counting to estimate the FLOPs for a given architecture. The input tensor is of shape `(B, S, D)` throughout.

[38] The notation surrounding floating-point operations is very confusing because another quantity of interest is the number of floating-point operations a given implementation can use *per-second*. Sometimes, people use FLOPS or FLOP/s to indicate the rate, rather than the gross-count which has the lower case "s", FLOPs, but there's little consistency in general. We will use FLOPs and FLOP/s.

[39] Since their FLOPs counts only scales as $\sim \mathcal{O}(n)$ where the omitted constant may be relatively large, but still negligible when all dimensions are big.

Essentials

The number of FLOPs to push a batch of B of sequence-length S examples through the forwards-pass of a decoder-only transformer is approximately $2BSN_{\text{params}}$ where the number of parameters accounts for any reductions due to tensor- and sequence-parallelism[a]. The backwards-pass costs about twice as much as the forwards-pass. This is true as long as $S \lesssim D$).

[a] A quick argument: a computation of the form $T_{a_0 \dots a_n j} = V_{a_0 \dots a_A i} M_{ij}$ requires $2A_0 \dots A_n IJ$ FLOPs where the capital letters represent the size of their similarly-index dimensions. Thus, the FLOPs essentially count the size of the matrix M (that is, IJ), up to a factor of 2 times all of the dimensions in V which weren't summed over. Therefore, passing a `(B, S, D)`-shaped tensor through the Transformer architecture would give $\sim 2BS\times$(sum of sizes of all weight-matrices) FLOPs, and that this last factor is also approximately the number of parameters in the model (since that count is dominated by weights). Thus, FLOPs $\approx 2BSN_{\text{params}}$. This is the correct as long as the self-attention FLOPs with $\mathcal{O}\left(S^2\right)$-dependence which we didn't account for here are actually negligible (true for $S \lesssim 10D$).

7.1 No Recomputation

Start with the case where there is no recomputation activations. These are the **model FLOPs** of [5], as compared to the **hardware FLOPs** which account for gradient checkpointing.

`CausalAttention`: Forwards The FLOPs costs:

- Generating the query, key, and value vectors: $6BSD^2$
- Attention scores: $2BDS^2$
- Re-weighting values: $2BDS^2$
- Final projection: $2BSD^2$

`MLP`: Forwards Passing a through the `MLP` layer, the FLOPs due to the first and second matrix-multiplies are equal, with total matrix-multiply FLOPs $4BSED^2$.

Backwards Pass: Approximate The usual rule of thumb is to estimate the backwards pass as costing twice the flops as the forwards pass. This estimate comes from just counting the number of $\mathcal{O}\left(n^2\right)$ matrix-multiply-like operations and seeing that for every one matrix multiplication that was needed in the forward pass, we have roughly twice as many similar operations in the backwards pass.

The argument: consider a typical sub-computation in a neural network which is of the form $z' = \phi\left(W \cdot z\right)$ where z', a are intermediate representations z, z', ϕ is some non-linearity, and where the matrix multiply inside the activation function dominates the forwards-pass FLOPs count, as above. Then, in the backwards pass for this sub-computation, imagine we are handed the upstream derivative $\partial_{z'}\mathcal{L}$. In order to complete backpropagation, we need both to compute $\partial_W \mathcal{L}$ to update W and also $\partial_z \mathcal{L}$ to continue backpropagation to the next layer down. Each of these operations will cost about as many FLOPs as the forwards-pass, hence the estimated factor of two (but, as we will see, this is a very rough estimate).

Being more precise, let z be `(D0, ... , Dn, J)`-shaped and let W be `(I, J)`-shaped such that it acts on the last index of z, making z' `(D0, ... , Dn, I)`-shaped. Denoting $D = \prod_i D_i$ be the number of elements along the D_i directions for brevity, the forward-FLOPs cost of the sub-computation is therefore $2DIJ$.

Adding indices, the two derivatives we need are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial W_{ij}} &= \frac{\partial \mathcal{L}}{\partial z'_{d_0 \dots d_n i}} \phi' \left((W \cdot z)_{d_0 \dots d_n i} \right) z_{d_0 \dots d_n j} \\ \frac{\partial \mathcal{L}}{\partial z_{d_0 \dots d_n j}} &= \frac{\partial \mathcal{L}}{\partial z'_{d_0 \dots d_n i}} \phi' \left((W \cdot z)_{d_0 \dots d_n i} \right) W_{ij} \ , \end{aligned} \tag{7.1}$$

which have shapes `(I, J)` and `(D0, ..., Dn, J)`, respectively. On the right side, z and $W \cdot z$ are cached and the element-wise computation of $\phi'(W \cdot z)$ has negligible FLOPs count, as discussed above: its contribution is $\mathcal{O}(1/I)$ suppressed relative to the matrix-multiplies. The FLOPs count is instead dominated by the broadcast-multiplies, sums, and matrix-products.

The two derivatives in (7.1) each have the same first two factors in common, and it takes DI FLOPs to multiply out these two `(D0, ... , Dn, J)`-shaped tensors into another result with the same shape. This contribution is again $\mathcal{O}(1/I)$ suppressed and hence negligible. Multiplying this factor with either $z_{d_0 \dots d_n i}$ or W_{ij} and summing over the appropriate indices requires $2DIJ$ FLOPs for either operation, bringing the total FLOPs to $4DIJ$, which is double the FLOPs for this same sub-computation in the forward-direction, hence the rough rule of thumb[40].

Backwards Pass: More Precise TODO

Total Model FLOPs The grand sum is then[41]:

$$C^{\text{model}} \approx 12BDLS\left(S + (2+E)\,D\right) \ . \tag{7.2}$$

We can also phrase the FLOPs in terms of the number of parameters (10.10) as

$$C^{\text{model}}\big|_{T=1} = 6BSN_{\text{params}} \times \left(1 + \mathcal{O}\left(S/D\right)\right) \tag{7.3}$$

where we took the $T = 1, D \gg S$ limit for simplicity and we note that BS is the number of total tokens in the processed batches.

8 Training Time

Training is generally compute bound (see App. D) and based on the results of Sec. 7 the quickest one could possibly push a batch of data through the model is

$$t_{\text{min}} = \frac{C^{\text{model}}}{\lambda_{\text{FLOP/s}}} \ . \tag{8.1}$$

[40] Note also that the very first layer does not need to perform the second term in (7.1), since we do not need to backpropagate to the inputs, so the total backwards flops is more precisely $4DIJ(L-1) + 2DIJ$.

[41] With a large vocabulary, the cost of the final language model head matrix multiply can also be significant, but we have omitted its L-independent, $2BDSV$ contribution here.

Expanding to the entire training run, then with perfect utilization training will take a time

$$t_{\text{total}} \approx \frac{6 N_{\text{params}} N_{\text{tokens}}}{\lambda_{\text{FLOP/s}}} \ . \tag{8.2}$$

Adjust $\lambda_{\text{FLOP/s}}$ to the actual achievable FLOP/s in your setup to get a realistic estimate.

How many tokens should a model of size N_{params}? Scaling laws (Sec. 9) provide the best known answer, and the Summer 2023 best-guess is that we optimally have $N_{\text{tokens}} \approx 20 N_{\text{params}}$. So that the above is

$$t_{\text{total}} \approx \frac{120 N_{\text{params}}^2}{\lambda_{\text{FLOP/s}}} \ , \tag{8.3}$$

leading to quadratic growth in training time.

Note that the above is only correct if we are actually only spending C^{model} compute per iteration. This is not correct if we use gradient checkpointing and recomputation, in which case we alternatively spend true compute $C^{\text{hardware}} > C^{\text{model}}$, a distinction between **hardware FLOPs** and **model FLOPs**. Two corresponding efficiency measures are **model FLOPs utilization** (MFU) and **hardware FLOPs utilization** (HFU). If our iterations take actual time t_{iter}, then these are given by

$$\text{MFU} = \frac{t_{\text{iter}}}{t_{\text{min}}^{\text{model}}} \ , \quad \text{HFU} = \frac{t_{\text{iter}}}{t_{\text{min}}^{\text{hardware}}} \ , \tag{8.4}$$

where $t_{\text{min}}^{\text{model}}$ is (8.1) and $t_{\text{min}}^{\text{hardware}}$ is similar but using C^{hardware}.

9 Scaling Laws

Empirically-discovered scaling laws have driven the race towards larger and larger models.

Essentials

Decoder-only model performance improves predictably as a function of the model size, dataset size, and the total amount of compute. As of Summer 2023, there is little sign of hitting any kind of wall with respect to such scaling improvements.

The central parameters are:

- The number of non-embedding model parameters, as excising embedding params was found to generate cleaner scaling laws. Because our N_{params} has already been typically neglecting these parameters, we will just use this symbol in scaling laws and keep the above understanding implicit.[42] [23].
- C: total compute, often in units like PFLOP/s-days $\sim 10^{20}$ FLOPs
- N_{tokens}: dataset-size in tokens

[42] Presumably, the scaling laws are cleaner with these neglected because these params do not contribute directly to FLOPs, unlike most other parameters.

- $\mathcal{L}$: cross-entropy loss in nats

The specific form of any given scaling law should also be understood to apply to a pretty narrowly defined training procedure, in which choices like the optimizer, learning-rate scheduler, hyperparameter search budget, vocabulary size, tokenization, etc. are often rigidly set. Changing different components of the training procedure is liable to create different scaling laws (though nice laws of some form are still expected to exist).

9.1 Original Scaling Laws

The first scaling-laws were reported in [23]. Their simplest form relates the value of the cross-entropy loss *at convergence* (and in nats), $\mathcal{L}$, to the number of non-embedding parameter, dataset size in token, and the amount of compute, *in the limit* where only one of this factors is bottlenecking the model[43]. The laws (in our notation):

- $\mathcal{L}(N_{\text{params}}) \approx \left(N^{\star}_{\text{params}}/N_{\text{params}}\right)^{\alpha_N}$, with $\alpha_N \approx 0.076$ and $N^{\star}_{\text{params}} \approx 8.8 \times 10^{13}$
- $\mathcal{L}(N_{\text{tokens}}) \approx (N^{\star}_{\text{tokens}}/N_{\text{tokens}})^{\alpha_T}$, with $\alpha_T \approx 0.095$ and $N^{\star}_{\text{tokens}} \approx 5.4 \times 10^{13}$
- $\mathcal{L}(C) \approx (C^{\star}/C)^{\alpha_C}$, with $\alpha_C \approx 0.050$ and $C^{\star} \approx 3.1 \times 10^{8}$ PFLOP/s-days, where the batch size was assumed to be chosen to be compute optimal per the criteria they outline

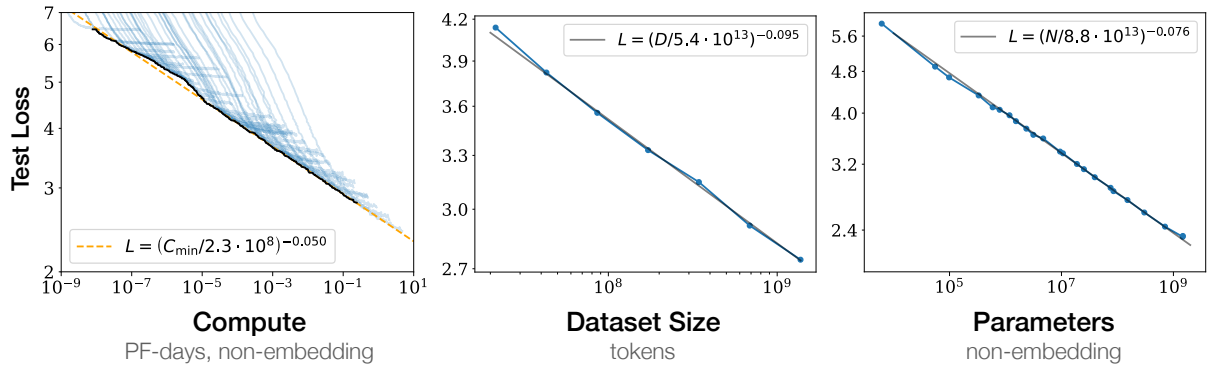

Figure 2. Original scaling laws from [23].

9.2 Chinchilla Scaling Laws

As of Summer 2023, the Chinchilla scaling laws in [24] are the de facto best scaling laws for guiding training. The central difference between [24] and [23] is that in the former they adjust their cosine learning-rate schedule to reflect the amount of planned training, while in the latter they do not[44].

[43] Unclear to me how you know when this is the case?

[44] The learning-rate schedule consist of a linear warm-up stage from a very small η up to the largest value η_{max}, after which the cosine bit kicks in: $\eta(s) = \eta_{\text{min}} + (\eta_{\text{max}} - \eta_{\text{min}}) \times \cos\left(\frac{\pi s}{2 s_{\text{max}}}\right)$ with s the step number. In Fig. A1 of [24] they demonstrate that having the planned s_{max} duration of the scheduler be longer than the actual number of training steps is detrimental to training (they do not study the opposite regime), which is effectively what was done in [23]. Probably the more important general point is again that the precise form of these scaling laws depend on details of fairly arbitrary training procedure choices, such as the choice of learning-rate scheduler.

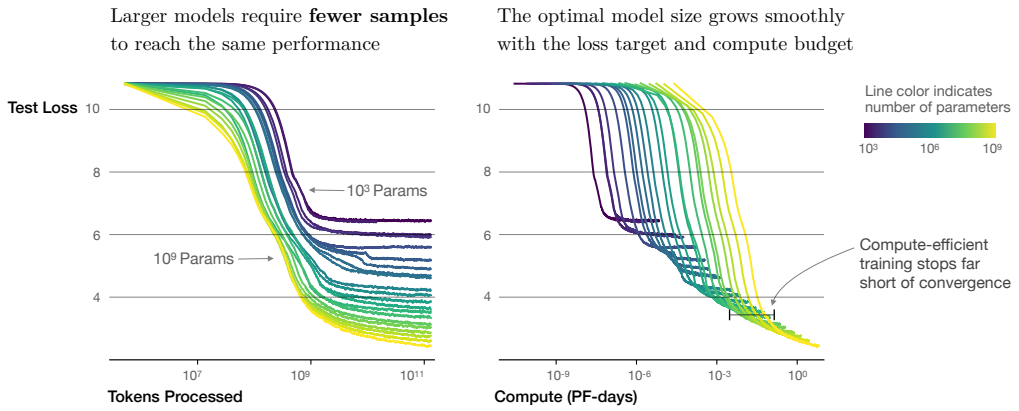

Figure 3. From [23]. Larger models are much more sample-efficient (faster).

Several different analyses are performed which all give very similar results. The outputs are the optimal values of $N_{\text{params}}, N_{\text{tokens}}$ given a compute budget C.

- They fix various buckets of model sizes and train for varying lengths. In their resulting loss-vs-FLOPs plot, they determine the model size which led to the best loss at each given FLOPs value, thereby generating and optimal model size vs compute relation.
- They fix various buckets of FLOPs budget and train models of different sizes with that budget, finding the optimal model size in each case. A line can then be fit to the optimal settings across FLOPs budgets in both the parameter-compute and tokens-compute planes.
- They perform a parametric fit to the loss[45]:

$$\mathcal{L}(N_{\text{params}}, N_{\text{tokens}}) = E + \frac{A}{N_{\text{params}}^{\alpha}} + \frac{B}{N_{\text{tokens}}^{\beta}} \ , \tag{9.1}$$

  fit over a large range of parameter and token choices. The best-fit values are:

$$E = 1.69 \ , \quad A = 406.4 \ , \quad B = 410.7 \ , \quad \alpha = 0.34 \ , \quad \beta = 0.28 \ . \tag{9.2}$$

  Using $C \approx 6 N_{\text{params}} N_{\text{tokens}}$, the above can be minimized at fixed compute either for number of parameter or the size of the dataset.

In all cases, the findings are that at optimality $N_{\text{params}} \sim N_{\text{tokens}} \sim C^{.5}$: both the parameter and tokens budget should be scaled in equal measure.

[45] In [24] they model the scaling of the test loss, while in [23] they use the training loss.

Part IV
Fine Tuning

10 Instruction Fine Tuning

Generally, instruction fine-tuning is a follow-on step after model pre-training[46]. The pre-training, pure next-token prediction task is altered to optimize an objective which now incorporates other data, typically information regarding human preferences[47].

10.1 Direct Preference Optimization

Direct Preference Optimization (DPO) [25] is a vast simplification of previous reinforcement-learning based methods (namely PPO-based ones [26]).

DPO aims to solve the RLHF optimization problem defined over a dataset $\mathcal{D} \sim (x, y_l, y_w)$ corresponding to prefixes (x) and pairs of preferred and dispreferred completions[48] (y_l, y_w). The relevant components are:

1. A baseline language model: $\pi_{\text{ref}}(y|x)$, usually a supervised fine-tuned model trained on high-quality data.

2. The to-be-trained model: $\pi_\theta(y|x)$, usually initialized to $\pi_{\text{ref}}(y|x)$. This is the *policy* in the literature.

3. A reward model which produces $p(y_w \succ y_l|x)$, the probability[49] y_w is favored over y_l. The reward function $r(x, y)$ reflects how well y completes the prefix x, in this context, and we assume the probability can be expressed in terms of the reward function $p(y_w \succ y_l|x) = p(r(x, y_w), r(x, y_l))$. The reward model is commonly an LLM with a scalar output head attached.

First, a quick review of RLHF, which proceeds in stages. First, $\mathcal{D}$ is used to train a reward model informed by the dataset $\mathcal{D}$. The optimal reward model $r_\star$ minimizes the binary cross-entropy loss over $\mathcal{D}$, which is just

$$\mathcal{L}_r = -E_{x,y_l,y_w\sim\mathcal{D}} \ln p(y_w \succ y_l|x) \ . \tag{10.1}$$

The reward model embodies human preferences and we want to transfer this knowledge to the language model π_θ. This can be done by optimizing π_θ to generate completions of inputs that lead

[46] A terminology note: pre-training is standard next-token training on an enormous, general dataset, supervised fine-tuning typically indicates additional, subsequent training on a higher-quality, maybe domain-specific dataset, and instruction fine-tuning follows.

[47] One failure mode this corrects for: next-token training would do best by replicating common mistakes in grammar or statements of fact which can be corrected for using these methods.

[48] I guess the l, w subscripts are for "lose" and "win"?

[49] Whether one completion is preferred over another is a probabalistic question since, e.g., not everyone in the population will agree.

to large rewards, reflecting human-preferred generations. In order to also keep the model from straying too far from its reference base, a tunable KL-divergence penalty is also added[50]:

$$\mathcal{L}_{\text{RLHF}} = E_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left(-r_\star(x,y) + \beta D_{\text{KL}}\left(\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)\right)\right) \ . \tag{10.2}$$

Reinforcement-learning methods are typically used to optimize the π_θ model and the generation step is particularly costly. In particular, the usual gradient-based optimization methods cannot be used because the loss depends on generated tokens which are discontinuous (non-differentiable) functions of the model's parameters.

DPO improves upon RLHF by skipping any generation step, removing the explicit reward function, and making the optimization problem amenable to gradient based methods by choosing a specific functional relation between the reward function $r(x,y)$ and the preference probability $p(y_w \succ y_l|x)$. Whereas RLHF minimizes the loss $\mathcal{L}_{\text{rlhf}}$ (10.2) subject to a fixed, optimal reward function found by first minimizing the reward loss $\mathcal{L}_r$ (10.1), DPO is essentially derived in the opposite direction: first, find the functional form of π_θ which minimizes the RLHF loss for an arbitrary reward function, and then use this form when minimizing of the cross-entropy defining the reward function[51].

The π_θ which minimizes the RLHF loss (10.2) for an arbitrary reward function $r(x,y)$ is given by[52]

$$\pi_\theta(y|x) = \frac{\pi_{\text{ref}}(y|x)e^{r(x,y)/\beta}}{Z(x)} \ , \tag{10.3}$$

where $Z(x) = \int \mathrm{d}y\, \pi_{\text{ref}}(y|x)e^{r(x,y)/\beta}$ is a intractable normalization (partition function) factor. However, if $p(y_w \succ y_l|x)$ only depends on $r(x,y_w)$ and $r(x,y_l)$ through their difference[53], these factors cancel out. Letting $p(y_w \succ y_l|x) = \sigma(r(x,y_w) - r(x,y_l))$, for some[54] σ, and eliminating the reward function in the cross-entropy loss via (10.3) reduces $\mathcal{L}_r$ to

$$\mathcal{L}_{\text{DPO}} = -E_{x,y_l,y_w\sim\mathcal{D}} \ln\sigma\left(\beta\left(\ln\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \ln\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right) \ , \tag{10.4}$$

which we've now renamed the DPO loss. The loss (10.4) can now be minimized by standard, gradient based methods without any generation step.

10.2 KTO: Preference Finetuning without Pairs

DPO requires a dataset of triplets: a prefix, one preferred completion, and one dispreferred completion. KTO alignment [27] attempts to reduce the inputs a prefix, a completion, and a binary signal indicating whether the output is desirable or not, since such datasets are easier to construct.

[50] We've written the above as a loss so that we're minimizing everywhere.

[51] This is analogous to minimizing the regular function $f(x,y)$ subject to also minimizing $g(x)$. This can either be done by solving the second for $x_\star$ and minimizing $f(x_\star,y)$ (the RLHF strategy), or first solving $\frac{\partial f}{\partial y} = 0$ to find $x_\star(y)$ and then minimizing $g(x_\star(y))$ (the DPO strategy).

[52] This is easy to show using the calculus of variations, though it's not the route taken in the paper. The explicit RLHF loss is $\mathcal{L}_{\text{RLHF}} = \int \mathrm{d}x\,\mathrm{d}y\, p(x)\pi_\theta(y|x)\left(-r(x,y) + \beta\ln\pi_\theta(y|x)\ /\pi_{\text{ref}}(y|x)\right)$ and we want to minimize this subject to the constraint that $\pi_\theta(y|x)$ is properly normalized. So, we use a Lagrange multiplier and extremize $\mathcal{L}' = \mathcal{L}_{\text{RLHF}} + \int \mathrm{d}x\,\mathrm{d}y\,\lambda(x)\pi_\theta(y|x)$. Solving $\frac{\delta\mathcal{L}'}{\delta\pi_\theta(y|x)} = 0$ yields (10.3).

[53] In [25], the DPO symmetry $r(x,y) \longrightarrow r(x,y) + f(x)$, for arbitrary $f(x)$, is said to induce an equivalence class relation between different reward functions.

[54] In the specific case where σ is the sigmoid function, this is known as the Bradley-Terry model.

The method is based on the ideas of Kahneman and Tversky and the central ingredient is a value function which monotonically maps outcomes to perceived values $v : \mathcal{Z} \longrightarrow \mathbb{R}$, with $\mathcal{Z}$ the space of outcomes. Some normalization point z_0 defines the boundary between positive and negative outcomes, the value function[55] is taken to be a function of $z - z_0$, and human value functions are known to be convex for $z > z_0$ (diminishing returns) and exhibit loss aversion[56].

KTO applies this framework to the usual text-prediction problem as in the following. The space of outcomes $\mathcal{Z}$ is the reward function value taken to be

$$r_\theta(x, y) \equiv \ln \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \ , \tag{10.5}$$

the difference in reference and model surprisal, as inspired by DPO. The reference point is just the expected value of the reward function over prefixes and trainable-model-generated completions, i.e., the KL divergence averaged over prefixes:

$$z_0 \equiv E_{y\sim\pi_\theta(y|x),x\sim D} r_\theta(x, y) = E_{x\sim D} D_{\text{KL}}(\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)) \ . \tag{10.6}$$

Splitting the space of completions into desirable and undesirable ones, $\mathcal{Y} = \mathcal{Y}_D \cup \mathcal{Y}_U$, the KTO loss[57] is taken to be:

$$\begin{aligned} \mathcal{L}_{\text{KTO}} &= -E_{x,y\sim D} v(r_\theta(x, y) - z_0) \\ v(r_\theta(x, y) - z_0) &\equiv \begin{cases} \lambda_D \sigma\left(\beta\left(r_\theta(x, y) - z_0\right)\right) & y \in \mathcal{Y}_D \\ \lambda_U \sigma\left(\beta\left(-r_\theta(x, y) + z_0\right)\right) & y \in \mathcal{Y}_U \end{cases} \end{aligned}$$

for hyperparameters[58] $\beta, \lambda_D, \lambda_U \in \mathbb{R}^+$ and where σ is the sigmoid function. So, $v(r_\theta(x, y) - z_0)$ is maximized by sending $r_\theta \longrightarrow \infty$ for desirable results and to $-\infty$ for undesirable ones, while the normalization point z_0 concentrates updates on examples whose rewards do not stray wildly from the average reward, which implicitly carries information about the reference model.

The reference point z_0 (10.6) is a problem, because it requires generation which is both expensive and not differentiable (the problem DPO solves). So, the authors perform a rough estimate of the scale and do not backpropagate through z_0, (which is a bit questionable).

[55] Which can be taken to satisfy $v(0) = 0$.

[56] Which I suppose means that $v\left(z - z_0\right) + v\left(z_0 - z\right) \leq 0$ for $z > 0$.

[57] They also add a constant term to the loss for normalization purposes which we have omitted. The KTO loss falls into the broader category of Human Aware Loss Objectives (HALOs) which are a general class of objectives that roughly fit into the Kahneman-Tversky form. See the paper for a further discussion and comparison of HALO vs non-HALO methods.

[58] Risk aversion would seem to require $\lambda_U > \lambda_D$, but the KTO paper empirically finds that the opposite regime performs better.

Part V
Parallelism

The simplicity of the Transformers architecture lends itself to a deep variety of parallelism strategies. We review some of them below.

10.3 Tensor Parallelism

Side Note:

I wrote a blog post on this here.

In **Tensor Parallelism**, sometimes also called **Model Parallelism**, individual weight matrices are split across devices [28]. We consider the `MLP` and `CausalAttention` layers in turn. Assume T-way parallelism such that we split some hidden dimension into T-equal parts across T workers[59]

Essentials

The cost of large weights can be amortized by first sharding its output dimension, resulting in differing activations across group members. Later, the activations are brought back in sync via a `AllReduce`. Weights which act on the sharded-activations can also be sharded in their input dimension. In the backwards pass, another `AllReduce` is required.

MLP It is straightforward to find the reasonable ways in which the weights can be partitioned. We suppress all indices apart from those of the hidden dimension for clarity.

The first matrix multiply $z_d W^0_{de}$ is naturally partitioned across the output index, which spans the expanded hidden dimension $e \in \{0, \ldots, ED-1\}$. This functions by splitting the weight matrix across its output indices across T devices: $W^0_{de} = W^0_{d(ft)} \equiv \bar{W}^0_{df\bar{t}}$ (again in `einops`-like notation, with bars denoting that the tensor and particular indices are sharded; see App. A), where in the split weights $\bar{t} \in \{0, \ldots, T-1\}$, and $f \in \left\{0, \ldots, \frac{ED}{T} - 1\right\}$. Each of the T workers compute one shard of $z_d \bar{W}^0_{df\bar{t}}$, i.e. each has a different value of $\bar{t}$.

Let the partial outputs from the previous step be $\bar{z}_{f\bar{t}}$ (batch-index suppressed), which are `(B, S, E*D/T, T)`-shaped, with the final dimension sharded across workers. The non-linearity ϕ acts element wise, and using the updated $\bar{z}_{f\bar{t}}$ to compute the second matrix multiply requires a splitting the weights as in $W^1_{ed'} = W^1_{(ft)d'} \equiv \bar{W}^1_{f\bar{t}d'}$ (dividing up the incoming e dimension), such that the desired output is computed as in $\bar{z}_{f\bar{t}} \cdot \bar{W}^1_{f\bar{t}d'}$, sum over $\bar{t}$ implied. Each device has only $\bar{t}$ component in the sum (a `(B, S, D)`-shaped tensor) and an `AllReduce` is used to give all workers the final result. This `AllReduce` is the only forward-pass collective communication[60].

[59] All T workers work on processing the same batch collectively. With $N > T$ workers, with N perfectly divisible by T, there are N/T different data parallel groups. Critical-path TP communications occur within each data parallel group and gradients are synced across groups. Ideally, all the workers in a group reside on the same node, hence the usual $T = 8$.

[60] The amount of communicated data is $\mathcal{O}(BSD)$.

One-line summary of the parallel decomposition:

$$z_{sd'} \leftarrow \phi\left(z_d W^0_{de}\right) W^1_{ed'} = \phi\left(z_d \bar{W}^0_{df\bar{t}}\right) \bar{W}^1_{f\bar{t}d'} \ . \tag{10.7}$$

The progression of tensor shapes held by any single worker is

1. `(B, S, D)`
2. `(B, S, E*D/T)`
3. `(B, S, D)`

In the backwards pass, another `AllReduce` (see App. B) is needed for proper gradient computations with respect to the first `Linear` layer's outputs. This is true whenever an operation producing a sharded output involved non-sharded tensors: if an operation $\bar{y}_{\bar{r}} = F(x, \ldots)$ produces a sharded output from an unsharded in put x (all other indices suppressed), the derivative with respect to x requires a sum over ranks, $\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial \bar{y}_{\bar{r}}} \frac{\partial \bar{y}_{\bar{y}}}{\partial x}$. Note that each worker will have to store all components of the input z for the backward pass.

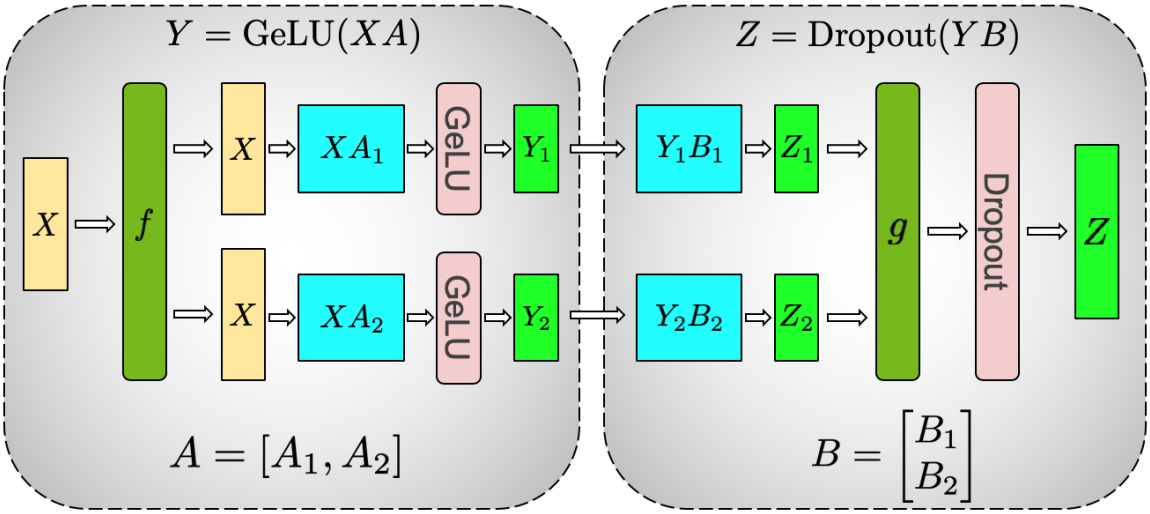

Figure 4. Tensor parallelism for the `MLP` layers. Graphic from [28]. The f/g operations are the collective identity/`AllReduce` operations in the forwards pass and the `AllReduce`/identity operations in the backwards pass.

Attention Because the individual attention head computations are independent, they can be partitioned across T workers without collectively communications. An `AllReduce` is needed for the final projection, however, which results in the various re-weighted values y_{bsea} (1.5).

To review, the attention outputs z'_{sd} generated from inputs z_{sd} can be expressed as

$$z'_{sea} = \mathrm{MHA}(q_{sea}, k_{sea}, v_{sea}) O_{ead} \tag{10.8}$$

where:

- We have split the d-index as in $z_{sd} \longrightarrow z_{s(ea)}$ with e and a the head-dimension and head-index
- $q_{sea}, k_{sea}, v_{sea}$ are the query, keys and values derived from the inputs
- MHA is the multi-head attention function, whose outputs are the same shape as its value inputs
- The dual sum over head-dimension index (e) and attention-head-index (a) is the sum-and-concatenate step from the more explicit description in Sec. 1.3
- `Dropout` and biases were ignored for simplicity

In order to parallelize the above T-ways, we simply shard across the dimension a which indexes the different attention heads. The MHA computations all process in embarassingly-parallel fashion, and an all-reduce is needed to complete the sum over the a-index across devices.

The collective communications story is essentially equivalent to that of the `MLP` layers[61]: one `AllReduce` is needed in the forwards pass and one `AllReduce` in the backwards-pass.

The progression of tensor shapes held by any single worker is

1. `(B, S, D)`
2. `(B, S, D/A, A/T)`
3. `(B, S, D)`

It is worth comparing the communications and FLOPs costs of these sharded layers. Each layer costs $\mathcal{O}\left(BS\left(4+2E\right)D^2/T\right)$ FLOPs and communicates $\mathcal{O}\left(BSD\right)$ bytes and so the communication-to-compute-time ratio is

$$\frac{t_{\text{compute}}}{t_{\text{comms}}} \sim \frac{\left(4+2E\right)D}{T} \times \frac{\lambda_{\text{comms}}}{\lambda_{\text{FLOP/s}}} \,. \tag{10.9}$$

Since[62] $\frac{\lambda_{\text{comms}}}{\lambda_{\text{FLOP/s}}} \sim 10^{-3}$FLOPs/B, communication and compute take similar times when $D \sim \mathcal{O}\left(10^3\right)$ for typical setups with $T \sim \mathcal{O}\left(10\right)$ and so tensor-parallelism requires $D \gtrsim 10^4$ to reach similar efficiency to the non-tensor-parallel implementations.

Embedding and LM Head Last, we can apply tensor parallelism to the language model head, which will also necessitate sharding the embedding layer, if the two share weights, as typical.

For the LM head, we shard the output dimension as should be now familiar, ending up with T different `(B, S, V/T)`-shaped tensors, one per group member. Rather than communicating these large tensors around and then computing the cross-entropy loss, it is more efficient to have each worker compute their own loss where possible and then communicate the scalar losses around[63].

For a weight-tied embedding layer, the former construction requires `AllReduce` in order for every worker to get the full continuous representation of the input.

[61] The amount of communicated data is again $\mathcal{O}\left(BSD\right)$.

[62] Assuming $\lambda_{\text{FLOP/s}}$ ∼100 TFLOP/s and $\lambda_{\text{comms}} \sim$ 100 GiB/s.

[63] In more detail, given the gold-answers y_{bs} for the next-token-targets, a given worker can compute their contribution to the loss whenever their `(B, S, V/T)`-shaped output $z_{bsv'}$ contains the vocabulary dimension v_* specified by y_{bs}, otherwise those tensor components are ignored.

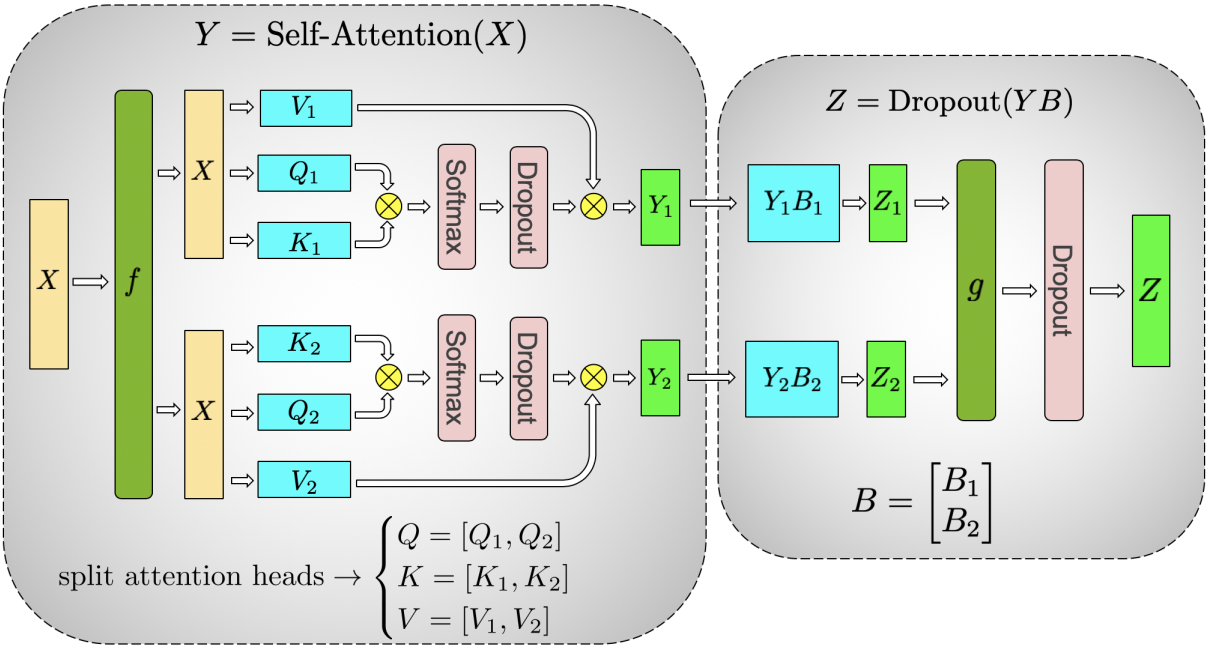

Figure 5. Tensor parallelism for the `CausalAttention` layers. Graphic from [28]. The f/g operators play the same role as in Fig. 4.

LayerNorm and Dropout `LayerNorm` instances are not sharded in pure tensor parallelism both because there is less gain in sharding them parameter-wise, but also sharding `LayerNorm` in particular would require additional cross-worker communication, which we wish to reduce as much as possible. `Dropout` layers are also not sharded in where possible in pure tensor parallelism, but sharding the post-attention `Dropout` layer is unavoidable. It is the goal of sequence parallelism is to shard these layers efficiently; see Sec. 10.4.

Effects on Memory The per-worker memory savings come from the sharding of the weights and the reduced activation memory from sharded intermediate representations.

The gradient and optimizer state memory cost is proportional to the number of parameters local to each worker (later we will also consider sharding these components to reduce redundantly-held information). The number of parameters per worker is reduced to

$$N_{\text{params}} \approx (4 + 2E)\frac{LD^2}{T} \ , \tag{10.10}$$

counting only the dominant contribution from weights which scale with L, since every weight is sharded. Since all non-activation contributions to training memory scale with N_{params}, this is a pure $1/T$ improvement.

The per-layer activation memory costs (6.5) and (6.6) are altered to:

$$M_{\text{act}}^{\texttt{Attention}} = BS\left(\left(p + \frac{4p}{T} + 1\right)D + \left(\frac{2p+1}{T}\right)AS\right)$$

$$M_{\text{act}}^{\texttt{MLP}} = \left(\frac{2Ep}{T} + p + 1\right) BDS \ . \tag{10.11}$$

The derivation is similar to before. Adding in the (unchanged) contributions from `LayerNorm` instances, the total, leading order activation memory sums to

$$M_{\text{act}}^{\text{total}} \approx 2BDLS\left(p\left(2 + \frac{E+2}{T}\right) + 1\right) + ABLS^2\left(\frac{2p+1}{T}\right) \ . \tag{10.12}$$

Again, the terms which did not receive the $1/T$ enhancement correspond to activations from unsharded `LayerNorm` and `Dropout` instances and the $1/T$'s improvements can be enacted by layering sequence parallelism on top (Sec. 10.4).

10.4 Sequence Parallelism

In (10.12), not every factor is reduced by T. **Sequence Parallelism** fixes that by noting that the remaining contributions, which essentially come from `Dropout` and `LayerNorm`[64], can be parallelized in the sequence dimension (as can the residual connections).

The collective communications change a bit. If we shard the tensors across the sequence dimension before the first `LayerNorm`, then we want the following:

1. The sequence dimension must be restored for the `CausalAttention` layer
2. The sequence should be re-split along the sequence dimension for the next `LayerNorm` instance
3. The sequence dimension should be restored for the `MLP` layer [65]

The easiest way to achieve the above is the following.

1. If the tensor parallelization degree is T, we also use sequence parallelization degree T.
2. The outputs of the first `LayerNorm` are `AllGather`-ed to form the full-dimension inputs to the `CausalAttention` layer
3. The tensor-parallel `CausalAttention` layer functions much like in Fig. 5 *except* that we do not re-form the outputs to full-dimensionality. Instead, before the `Dropout` layer, we `ReduceScatter` them from being hidden-sharded to sequence-sharded and pass them through the subsequent `Dropout`/`LayerNorm` combination, similar to the first step
4. The now-sequence-sharded tensors are reformed with another `AllGather` to be the full-dimensionality inputs to the `MLP` layer whose final outputs are similarly `ReduceScatter`-ed to be sequence-sharded and are recombined with the residual stream

The above allows the `Dropout` mask and `LayerNorm` weights to be sharded T-ways, but if we save the full inputs to the `CausalAttention` and `MLP` layers for the backwards pass, their contributions to the activation memory are not reduced (the p-dependent terms in (10.11)). In [5], they solve

[64] Recall, though, from Sec. 1.2 that the parameters in `LayerNorm` are completely redundant and can simply be removed without having any effect on the expressive capabilities of the architecture.

[65] This doesn't seem like a hard-requirement, but it's what is done in [5].

this by only saving a $1/T$ shard of these inputs on each device during the forward pass and then performing an extra `AllGather` when needed during the backwards pass. Schematics can be sen in Fig. 6 and Fig. 7 below. The activation memory is then reduced to:

$$M_{\text{act}}^{\text{total}} = \frac{2BDLS\left(p(E+4)+1\right)}{T} + \frac{ABLS^2\left(2p+1\right)}{T} + \mathcal{O}\left(BSV\right) \ . \tag{10.13}$$

In more detail:

- The norms are just linear operations on the z_{sd}, $z'_{sd} = \text{Norm}\left(z_{sd}\right)$, and so we split and shard cross the sequence dimension $z_{sd} \longrightarrow z_{(tr)d} \equiv \bar{z}_{\bar{t}rd}$ with the TP-index t sharded across devices.
- The residual stream is also sharded across the sequence dimension.
- The sharded outputs $\bar{z}_{\bar{t}rd}$ must be re-formed to create the attention and MLP inputs via an `AllGather`. There is an optimization choice here: either the re-formed tensors can be saved for the backward pass (negating the $1/T$ memory savings) or they can be re-formed via an `AllGather`, at the cost of extra communication.
- Both the MLP and attention layers need to produce final sums of the form $\bar{y}_{s\bar{y}e}\bar{O}_{\bar{t}ed}$ for some intermediate $\bar{y}$ and weight $\bar{O}$ sharded across the TP-dimension $\bar{t}$. The outputs are added to the sequence-sharded residual stream, and so sum is optimally computed through an `ReduceScatter` with final shape $\bar{z}_{\bar{t}'rd} = z_{(t'r)d} = z_{sd} = \bar{y}_{s\bar{t}e}\bar{O}_{\bar{t}ed}$. This `ReduceScatter` (along with the `AllGather` mentioned above) replace the `AllReduce`s from the tensor-parallel case and have the same overall communication cost.

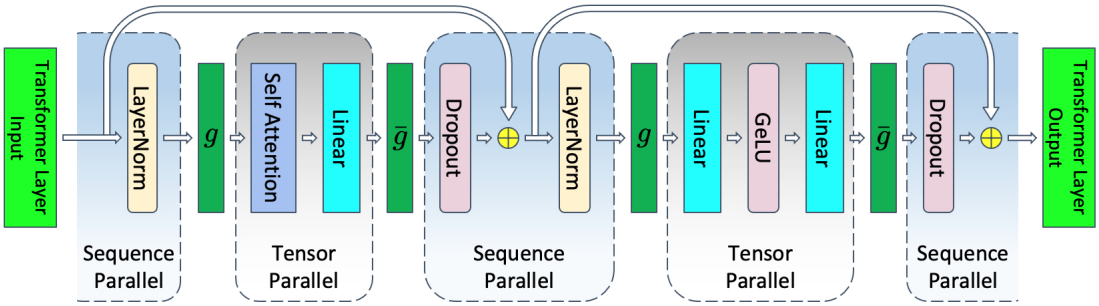

Figure 6. Interleaved sequence and tensor parallel sections. g and $\bar{g}$ are `AllGather` and `ReduceScatter` in the forward pass, respectively, and swap roles in the backwards pass. Graphic from [28].

10.5 Ring Attention

Ring Attention [29] is roughly a distributed version of Flash Attention 2.6: it enables extremely long sequence-length processing by never realizing the entire $\mathcal{O}\left(S^2\right)$ attention scores at once.

It works by sharding over the sequence dimension. Let z_{sd} is the (batch-dim suppressed) residual stream of a non-sharded Transformer[66]:

$$z_{sd} = \texttt{Softmax}_{s'}\left(q_{sd'}k_{s'd'}\right)v_{s'd} \ , \tag{10.14}$$

[66] Like in Sec. 2.6, we omit any normalization factor inside the `Softmax`.

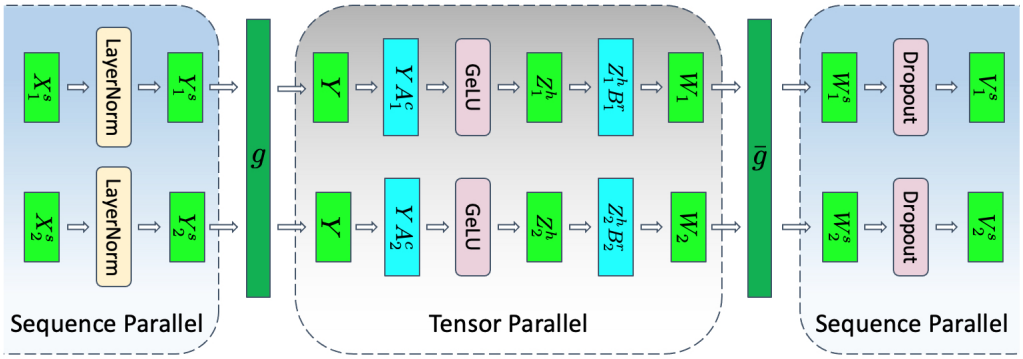

Figure 7. Detail of the sequence-tensor parallel transition for the `MLP` . Graphic from [28].

suppressing the causal mask for simplicity of presentation.

Then in Ring Attention, we shard over R devices via $z_{sd} \longrightarrow z_{\bar{r}td}$, and similar for other tensors, to compute the sharded outputs

$$\begin{aligned} z_{\bar{r}td} &= \texttt{Softmax}_{\bar{w}x}\left(q_{\bar{r}td'} k_{\bar{w}xd'}\right) v_{\bar{w}xd} \\ &= \frac{\exp\left(q_{\bar{r}td'} k_{\bar{w}xd'}\right)}{\sum_{\bar{w}'x'} \exp\left(q_{\bar{r}td''} k_{\bar{w}'x'd''}\right)} v_{\bar{w}xd} \\ &\equiv \frac{Z_{\bar{r}td}}{\sum_{\bar{w}'x'} \exp\left(q_{\bar{r}td''} k_{\bar{w}'x'd''}\right)} \\ &\equiv \frac{Z_{\bar{r}td}}{L_{\bar{r}t}} \end{aligned} \tag{10.15}$$

where we introduced some notation which will be useful blow. Ring Attention is essentially an algorithm for computing the sharded sums over barred indices via communication. Since the MLP layers act on every sequence position identically, only the Attention layers (and the loss computation) require special handling.

The algorithm performs the $\bar{w}$ sum as a loop. We present the simplified case without a causal mask or maximum attention score tracking. These are important omissions[67].

Algorithm 8 Ring Attention (Naive - Missing causal mask/max tracking.)

1: Initialize $Z_{\bar{r}td}$, $L_{\bar{r}t}$ to zeros
2: Populate the key, query, and value shards $q_{\bar{r}td}, k_{\bar{w}xd'}, v_{\bar{w}xd'}$ with $\bar{r} = \bar{w} = r$ on rank r
3: **for** $w \in \{r, \ldots, R-1, 0, \ldots, r-1\}$ **do** ▷ Computing components $z_{\bar{r}td}$ $\forall t, d$
4: **if** $w \neq (r-1) \mod R$ **then** prefetch shards $k_{(\bar{w}+1)xd}, v_{(\bar{w}+1)xd}$ $\forall x, d$
5: $Z_{\bar{r}td} \leftarrow Z_{\bar{r}td} + \exp\left(q_{\bar{r}td'} k_{\bar{w}xd'}\right) v_{\bar{w}xd}$ ▷ Can use flash attention kernels here
6: $L_{\bar{r}t} \leftarrow L_{\bar{r}t} + \sum_x \exp\left(q_{\bar{r}td'} k_{\bar{w}xd'}\right)$ ▷ Can use flash attention kernels here
7: $z_{\bar{r}td} \leftarrow \frac{Z_{\bar{r}td}}{L_{\bar{r}t}}$

[67] See [30] for causal mask efficiency considerations.

At every step in the loop in the algorithm we are computing the sums $\exp\left(q_{\bar{r}td'}k_{\bar{w}xd'}\right)v_{\bar{w}xd}$ and $\sum_x \exp\left(q_{\bar{r}td'}k_{\bar{w}xd'}\right)$ for fixed values of $\bar{r}, \bar{w}$ and all values of the other indices. These are precisely the ingredients that go into the usual attention computation and for this reason it's possible to use flash attention kernels for every individual step. `torch` implementations of Ring Attention which leverage flash attention kernels can be found here and here.

The full forms of the forwards and backwards passes are again similar to those of flash attention; see Sec. 2.6.1.

10.5.1 The Causal Mask

A naive, row-major sharding of the queries, keys, and vectors is highly suboptimal for causal attention because it leads to idling GPUs. Sharding the queries and keys as in $q_s = q_{(\bar{r}t)}$ and $k_{s'} = k_{(t'\bar{r}')}$ in row-major order[68], causality means that the entire chunked attention computation will be trivial for any iteration in which $r' > r$. This is the case for $R-1$ iterations for the $r = 0$ GPU, for instance.

So, we shouldn't shard this way for ring attention. In [30] they demonstrate the speed-up achieved by just reversing the sharding pattern to column-major: $q_s = q_{(t\bar{r})}$ and $k_{s'} = k_{(t'\bar{r}')}$ which guarantees non-trivial work for every GPU on every iteration, which they call striped ring attention. In the `ring-flash-attention` repo, they come up with yet another sharding strategy ("zig-zag" attention; see this github issue) which increases efficiency even more. Their strategy can't be naturally writtein in `einops` notation, but it is easy enough to explain: they split the sequence length into $2R$ sequential chunks and give zero-indexed chunks r and $2R-r-1$ to GPU r, which ends up optimally distributing the work.

We analyze the efficiency of each strategy now. Let q_s be sharded to $q_{\bar{r}t}$ according to the strategy's specifics and similar for k_s. On the first iteration every rank has keys and queries with identical sequence positions, meaning $\frac{\frac{S}{R}\left(\frac{S}{R}+1\right)}{2}$ positions will be attended to on every rank. The difference comes about in the subsequent iterations:

1. For naive ring attention, it's all or nothing. $q_{\bar{r}t}$ can attend to all of $k_{\bar{w}x}$ if $\bar{r} \geq \bar{w}$. This means that at least one rank needs to perform S^2/R^2 operations every iteration (after the first one), bottlenecking processes which have no work.

2. For striped attention ring attention, rank $r = R-1$ will have queries corresponding to positions $\{R-1, 2R-1, \ldots, S-1\}$ and it will always be able to attend to $\frac{\frac{S}{R}\left(\frac{S}{R}+1\right)}{2}$ positions at every iteration, just like on the first iteration. This rank (and others which perform the same number of operations for some iterations) is the bottleneck, since rank $r = 0$, which owns sequence positions $\{0, R, \ldots, S-R-1\}$ is only ever able to attend to $\frac{\frac{S}{R}\left(\frac{S}{R}-1\right)}{2}$ positions, since its zero-position component can never attend to anything after the first iteration. This mismatch between ranks is suboptimal.

3. For zig-zag attention, there are two scenarios. Each produces the same amount of work, which makes this strategy optimal. When $\bar{r} < \bar{w}$ the two sets of position indices covered

[68] That is, $s = \bar{r}T + t$ for $\bar{r} \in \{0, \ldots, R-1\}$ and $t \in \{0, \ldots T-1\}$ with $S = RT$.

by t on $q_{\bar{r}t}$ fall between those[69] covered by the x index on $k_{\bar{w}x}$. That is, every index in the query can attend to exactly half of the indices on the keys: $\frac{S^2}{2R^2}$ operations. In the opposite[70] $\bar{w} < \bar{r}$ case, the query positions sandwich those of the keys and so the upper half of the query positions can attend to all of the key's positions, again $\frac{S^2}{2R^2}$ operations. So work is perfectly distributed and no rank serves as a bottleneck.

10.6 Pipeline Parallelism

TODO

[69] Example: four ranks with sequence length $S = 8$, the rank-zero queries cover positions $\{0, 7\}$ while the rank-one keys cover $\{2, 6\}$, so the former sandwich the latter.

[70] The $\bar{w} = \bar{r}$ case was already treated in the first iteration.

Part VI
Vision

Notes on the usage of Transformers for vision tasks.

11 Vision Transformers

The original application of the Transformers architecture [31] divides 2D images into patches of size $P \times P$, e.g. flattening a three-channel i_{xyc} image to shape f_{sd} where $d \in \{0, \ldots, P^2C - 1\}$ and the effective sequence length runs over $s \in \{0, L^2C/P^2 - 1\}$, for an $L \times L$ sized image[71]. A linear projection converts the effective hidden dimension here to match match the model's hidden dimension. These are known as **Patch Embeddings**.

Since there is no notion of causality, no causal mask is needed. A special `[CLS]` token is prepended and used to generate the final representations z_{bd} for a batch of images. This can be used for classification, for instance, by adding a classification head. The original training objective was just that: standard classification tasks.

12 CLIP

CLIP (Contrastive Language-Image Pre-Training) [32] is a technique for generating semantically meaningful representations of images. The method is not necessarily Transformers specific, but the typical implementations are based on this architecture.

The core of CLIP is its training objective. The dataset consists of image-caption pairs (which are relatively easy to extract; a core motivation), the CLIP processes many such pairs and then tries to predict which images match to which captions. This is thought to inject more semantic meaning into the image embeddings as compared with, say, those generated from the standard classification task.

A typical implementation will use separate models for encoding the text and image inputs. The two outputs are t_{bd} and i_{bd} shaped[72], respectively, with batch and hidden dimensions, and are canonically trained so that the similarity score between any two elements is a function of their dot-product.

The original CLIP recipe:

1. Process the text bracketed with `[SOS]` and `[EOS]` insertions, use a normal Transformer architecture[73], and extract the last output from the `[EOS]` token as the text embedding: $i_{bd} = z_{bsd}\big|_{s=-1}$.
2. Process the image with a vision transformer network.

[71] Example: for a 256×256, three-channel image with a 16×16 patch size, the effective sequence length is 768.

[72] There may also be another linear projection from the actual model outputs to a common space, too. Obviously, this is also necessary if the hidden dimensions of the two models differ.

[73] The original CLIP paper keeps the causal mask.

3. Project to a common dimensionality space, if needed.

4. Compute the logits through cosine similarity: $\ell_{bb'} = i_{bd} t_{b'd} / |i_b||t_b|$. These are used to define both possible conditional probabilities[74]:

$$P(i_b|t_{b'}) = \frac{e^{\ell_{bb'}}}{\sum_b e^{\ell_{bb'}}} \ , \quad P(t_{b'}|i_b) = \frac{e^{\ell_{bb'}}}{\sum_{b'} e^{\ell_{bb'}}} \tag{12.1}$$

5. Compute the cross-entropy losses in both directions and average:

$$\mathcal{L} = \frac{1}{2B} \sum_b \left(\ln P\left(i_b|t_b\right) + \ln P\left(t_b|i_b\right) \right) \ . \tag{12.2}$$

They also add a temperature to the loss, which they also train.

Post-training, the CLIP models can be used in many ways:

1. Using the vision model as a general purpose feature extractor. This is how many vision-language models work: the CLIP image embeddings form part of the VLM inputs.

2. Classification works by comparing the logits for a given image across embedded sentences of the form `This is an image of a <CLASS HERE>.`

[74] They differ by what is summed over the in the denominator, i.e., which dimension the `Softmax` is over.

Part VII
Mixture of Experts

13 Basics

The $\mathcal{O}\left(D^2\right)$ FLOPs count due to MLP layers[75] is untenable past a given point: inference and training just take too long. Mixture of Experts[76] (MoE) models address this concern by splitting single MLP layer into a number of "expert" MLP layers and route a subset of the tokens to a subset of the experts." MoE is a lever for changing the relation between the per-token FLOPs count and the overall parameter count. Example: comparing a dense and a MoE model at similar parameter counts, the expert layer's intermediate dimension is reduced by $\mathcal{O}\left(N_{\text{ex}}\right)$ (the number of experts) and the FLOPs count is also reduced by this factor. Perhaps unsurprisingly, MoE experts outperform and train faster than their FLOPs equivalent dense models (at the cost of more engineering complexity and a higher memory burden).

The general form of the MoE layer output is

$$z'_{sd} = G_{se}(z_{sd}, \ldots) E_{esd}\left(z_{sd}\right) \tag{13.1}$$

where $G_{se}(z_{sd}, \ldots) \in \mathbb{R}^{S \times N_{\text{ex}}}$ is a gating (i.e., weighting) function and $E_{esd}\left(z_{sd}\right) \in \mathbb{R}^{N_{\text{ex}} \times S \times D}$ is the usual MLP operation performed by the e-th expert. Many of the entries G_{es} are zero in practice, and only the computations $E_{esd}\left(z_{sd}\right)$ corresponding to non-trivial gating values are performed, of course. Different MoE variants are essentially differentiated by the specific form of their weighting function.

14 Routing

Choosing which experts process which tokens is crucial, affecting both the downstream model and engineering (i.e. throughput) performance. There are two dominant schemes:

1. **Token Choice**: each token selects a fixed number of experts. G_{se} is sparse over the expert index; see (13.1).
2. **Expert Choice**: each expert selects a fixed number of tokens. G_{se} is sparse over the token index; see (13.1).

Layered on top of this choice are the details of the routing mechanisms.

14.1 Token Choice vs Expert Choice

Token and expert choice both introduce a tensor $W_{de} \in \mathbb{R}^{D \times N_{\text{ex}}}$ which is used to produce a score between each token and expert: $S_{se} = z_{sd} W_{de}$. In each case, we perform a `topk` computation and

[75] The $\mathcal{O}\left(S^2\right)$ scaling of the self-attention layers is also untenable, but MoE only addresses the MLP layers.

[76] The original MoE research came out of Google: see [33], [34] and related work by these authors. An excellent MoE paper with open-source everything is here [35].

output a weighted sum of expert outputs: the two methods just differ in the dimension over which the `topk` is performed.

For token choice, the gating function is:

$$G_{se}^{\text{token}}(z_{sd}, W) = \texttt{Softmax}_e\left(\texttt{topk}_e\left(z_{sd} \cdot W_{de}\right)\right) \ , \tag{14.1}$$

where this `topk` just sets all non-top-k entries to $-\infty$. G_{se} is sparse in its expert dimension and has Sk non-trivial elements. While every token will get routed to k experts with token choice routing, the per-expert load can be very unbalanced. Some token-choice implementations require setting a maximum tokens per expert limit which in turn defines the capacity factor c: $\texttt{maxtok} = c \times \frac{S}{N_{\text{ex}}}$. Tokens exceeding this limit are just not sent through the expert MLP at all (but remain in the residual stream, of course).

Expert choice just performs the `Softmax` and `topk` on the sequence dimension, instead. The gating function is

$$G_{se}^{\text{expert}}(z_{sd}, W) = \texttt{Softmax}_s\left(\texttt{topk}_s\left(z_{sd} \cdot W_{de}\right)\right) \ , \tag{14.2}$$

with `topk` acting as in the token choice case. G_{se} is sparse along the sequence dimension and has $N_{\text{ex}}k$ non-trivial elements. A (potential) disadvantage of expert choice is that some tokens may not be routed to any expert at all, but every expert is at least guaranteed an equal load. In this case, we effectively have $k = c \times \frac{S}{N_{\text{ex}}}$, with c the capacity factor above.

15 MegaBlocks

The MoE computation maps awkwardly to the typical GPU primitives. Ideally the expert computations in (13.1) are parallelized as much as possible, but batched matrix multiplies (the closet common primitive) enforces equal token counts per expert, which is overly restrictive.

MegaBlocks [36] introduces the proper sparse kernels to handle general MoE computations without the need to enforce any hard per-expert token limits or introduce unnecessary padding. They call their method dropless MoE (dMoE).

16 MoE Variants

A collections of other MoE architecture choices.

16.1 Shared Experts

Shared experts forces one particular expert to always be used, with the motivation of having the differentiated expert serve as a common pool of knowledge.

Part VIII
Inference

17 Basics and Problems

The essentials of decoder-only inference is that a given input sequence x_{bs} is turned into a probability distribution p_{bsv} over the vocabulary for what the next token might be. Text is then generated by sampling from p_{bsv} in some way, appending that value to x_{bs} to create a one-token-longer sequence, and then repeating until desired.

There are various problems that naive implementations of the above face:

- Repeated computation from processing the same tokens in the same order repeatedly, at least for some sub-slice of x_{bs}.
- Inherently sequential computation, rather than parallel
- Sub-optimal sampling strategies. Just choosing the most-probably token at each new step, does not guarantee the most-probable overall sequence, for instance.

18 Generation Strategies

A quick tour of generation strategies. A very readable blog post comparing strategies can be found here.

18.1 Greedy

The most obvious generation strategy is to take the final, `(B, S, V)`-shaped outputs z_{bsv} and just take the next token to be the most-probable one (for the final position in the sequence): `next_token = z[:, -1].argmax(dim=-1)`. A very minimal `generate` method is as below:

```python
class DecoderOnlyGreedy(DecoderOnly):
    def __init__(self, *args, **kwargs):
        super().__init__(*args, **kwargs)

    def generate(self, inputs, max_length):
        """
        Naive, minimal generation method. Assumes inputs are already tokenized. max_length can be
        longer than the block_size, but only up to block_size tokens can ever be included in the
        context.
        """
        self.eval()
        outputs = inputs.clone()
        while outputs.shape[1] < max_length:
            context = outputs[:, -self.block_size :]
            last_token_pred_logits = self(context)[:, -1]
            most_probable_token = last_token_pred_logits.argmax(dim=-1)[:, None]
            outputs = torch.cat([outputs, most_probable_token], dim=-1)
        return outputs
```

There are various important, practical considerations which are ignored in the above implementation, including:

- Since we are taking the prediction from the last (`-1`-indexed) element in each sequence, it is crucial that all padding is *left*-padding, so that these final elements are meaningful.
- Models will signal the end of generation by outputting tokenizer-specific codes, and generation must respect these.

See the `generate` method from the `transformers` library for more fully-featured code (which, correspondingly, is not always easy to follow).

18.2 Simple Sampling: Temperature, Top-k, and Top-p

The next-most-obvious strategy is to choose the next token by drawing from the probability distribution defined by the z_{bsv}. There are various refinements of this idea.

A one-parameter generalization of this strategy introduces a (physics-motivated) **Temperature** which just adjusts the scale of the logits:

```
next_token = torch.multinomial((z[:, -1] / temp).softmax(dim=-1), num_samples=1)
```

assuming `z` are the final logits. Larger temperature yields a larger variance in the chosen tokens.

With temperature sampling, there is still a non-zero chance of choosing an extremely improbable token, which is undesirable if you do not trust the tails of the distribution. Two common truncation strategies which guard against this:

- **Top-k**: Only choose from the top-k most-probable examples (re-normalizing the probabilities across those k samples)
- **Top-p**: Only choose from the top-however-many most-probable examples whose probabilities sum to p (again re-normalizing probabilities). This is also sometimes called **nucleus sampling**.

18.3 Beam Search

Choosing, say, the most-probable next-token at each step is not guaranteed to yield the most probable *sequence* of tokens. So, **Beam Search** explores multiple sequences, using different branching strategies, and the probabilities of the various beam sequences can be compared at the end. Important note: generating the most-probable text is not necessarily equal to the most human-like text [37].

18.4 Speculative Decoding

Speculative decoding [38] is an excellent idea: use a cheaper "draft" model to perform the slow, iterative generation steps and check its work with the full model. Using a detailed-balance-like construction, it can be guaranteed that this speculative decoding generation strategy creates text drawn from the same distribution as the full model.

Informally, the algorithm is:

1. Generate γ tokens with the draft model, whose distribution is $q(x_t|x_{\text{prefix}})$, $t \in \{0, \ldots, \gamma - 1\}$. Write the generated tokens as z_t.

2. Pass the prefix and all γ generated tokens z_t through the full model, which computes probabilities via its distribution $p(x_t|x_{\text{prefix}})$.

3. For every generated token z_t, accept it unconditionally if $q(z_t|x_{\text{prefix}}) \leq p(x_t|z_{\text{prefix}})$. If $q(z_t|x_{\text{prefix}}) > p(z_t|x_{\text{prefix}})$, instead accept the token with only probability[77] $\frac{p}{q}$.

4. If only the first $n < \gamma$ tokens are accepted, generate token $n+1$ from a modified distribution $p'(x_t|x_{\text{prefix}}) = F\left[p(x), q(x)\right]$ built from the two model predictions and chosen (as will be shown) such that the entire algorithm generates the correct distribution. $n+1$ tokens are created in this case[78].

5. If all of the γ tokens are accepted, generate token $\gamma + 1$ from the full model's outputs.

Proof of correctness and the derivation of $p'(x)$: let $Q(x_t|x_{\text{prefix}})$ be the distribution described above. Then this can be broken down according to conditioning on the draft token and whether or not the draft token was accepted. Dropping the prefix condition for brevity and A and R stand for rejected and accepted, respectively, we have

$$\begin{aligned} Q(x_t) &= \sum_{z_t} Q(x_t|z_t, A)P(A|z_t)q(z_t) + Q(x_t|z_t, R)P(R|z_t)q(z_t) \\ &= \sum_{z_t} \delta_{x_t,z_t} \times \min\left(1, \frac{p(z_t)}{q(z_t)}\right) \times q(z_t) + p'(x_t)\left(1 - \min\left(1, \frac{p(z_t)}{q(z_t)}\right)\right) \times q(z_t) \\ &= \min\left(q(x_t), p(x_t)\right) + p'(x_t) \sum_{z_t} \left(1 - \min\left(1, \frac{p(z_t)}{q(z_t)}\right)\right) \times q(z_t) \end{aligned} \tag{18.1}$$

The sum is just some constant (denoted by $1-\beta$ in the paper, which should really have a t subscript) and so choosing

$$p'(x_t|x_{\text{prefix}}) \equiv \frac{p(x_t|x_{\text{prefix}}) - \min\left(q(x_t|x_{\text{prefix}}), p(x_t|x_{\text{prefix}})\right)}{1 - \beta} \tag{18.2}$$

achieves the goal of getting $Q(x_t|x_{\text{prefix}}) = p(x_t|x_{\text{prefix}})$. It can be verified that this distribution is properly normalized.

An approximate analysis for choosing the optimal value of γ can be found in the paper.

19 The Bare Minimum and the kv-Cache

There are two separate stages during generation. First, an original, to-be-continued series of prompts x_{bs} can be processed in parallel to both generate the first prediction and populate any intermediate values we may want to cache for later. We follow [39] and call this the **prefill** stage. For this procedure, we require the entire x_{bs} tensor.

[77] This choice is not fundamental; it just makes following expressions nicer.

[78] We cannot generate more tokens because drawing from p' effectively changes the prefix that the full model should use.

In the second, iterative part of generation (the **decode** stage) we have now appended one-or-more tokens to the sequence and we again want the next prediction, i.e. `z[:, -1, :]` for the last-layer outputs z_{bsd}. In this stage, we can avoid re-processing the entire x_{bs} tensor and get away with only processing the final, newly added token, *if* we are clever and cache old results (and accept a very reasonable approximation).

The important pieces occur in the `CausalAttention` layer, as that's the only location in which the sequence index is not completely parallelized across operations. Referring back to Sec. 1.3, given the input z_{bsd} of the `CausalAttention` layer, the re-weighted value vectors[79] $w^a_{bss'd}v^a_{bs'f}$ are the key objects which determine the next-token-prediction, which only depends on the $s = -1$ index values. Therefore, we can cut out many steps and the minimum requirements are:

- Only the attention weights $w^a_{bss'd}$ with $s = -1$ are needed
- The only query values q^a_{bsd} needed to get the above are those with $s = -1$
- Every component of the key and value vectors k^a_{bsd}, v^a_{bsd} is needed, but because of the causal mask, all components except for the last in the sequence dimension ($s \neq -1$) are the same as they were in the last iteration, up to a shift by one position[80]

So, we are led to the concept of the **kv-cache** in which we cache old key and query vectors for generation. The cache represents a tradeoff: fewer FLOPs are needed for inference, but the memory costs are potentially enormous, since the size of the cache grows with batch size and sequence length:

$$M_{\text{kv-cache}} = 2pBDLS/T \ , \tag{19.1}$$

in the general case with tensor-parallelism. This can easily be larger than the memory costs of the model parameter: $M^{\text{inference}}_{\text{params}} \sim pN_{\text{params}} \sim pLD^2$ (dropping $\mathcal{O}(1)$ factors), so that the cache takes up more memory when $BS \gtrsim D$, i.e. when the total number of token exceeds the hidden dimension. Using the kv-cache eliminates a would-be $\mathcal{O}\left(S^2\right)$ factor in the FLOPs needed to compute a new token, reducing it to linear-in-S dependence everywhere.

A very minimal implementation[81] is below:

```
class CausalAttentionWithCache(CausalAttention):
    def __init__(self, *args, **kwargs):
        super().__init__(*args, **kwargs)
        self.cached_keys = self.cached_values = None

    def forward(self, inputs, use_cache=True):
        """Forward method with optional cache. When use_cache == True, the output will have a
        sequence length of one."""
```

[79] Summed over s', but concatenating the different a values over the f dimension.

[80] This is where we need to accept a mild approximation, if using a sliding attention window. With an infinite context window, if we add a label t which indexes the iteration of generation we are on, then we would have that $z^{(t+1)}_{bsd} = z^{(t)}_{b(s-1)d}$ for every tensor in the network, except for when $s = -1$, the last position. The finiteness of the context window makes this statement slightly inaccurate because we can only ever keep K positions in context and the loss of the early tokens upon sliding the window over will slightly change the values in the residual stream.

[81] Warning: very non-optimized code! Purely pedagogical.

```
        if not use_cache:
            return super().forward(inputs)
        if self.cached_keys is None:
            # If the cache is not yet initialized, we need all q, k, v values.
            assert (
                self.cached_values is None
            ), "If cached_keys is None, cached_values should be None, too"
            queries, keys, values = self.get_qkv(inputs)
        else:
            # Otherwise, we only need q, k, v values for the last sequence position.
            queries, new_keys, new_values = self.get_qkv(inputs[:, [-1]])
            keys = [torch.cat([ck, nk], dim=1) for ck, nk in zip(self.cached_keys, new_keys)]
            values = [torch.cat([cv, nv], dim=1) for cv, nv in zip(self.cached_values, new_values)]
        # Update or initialize the cache.
        self.cached_keys = [k[:, -self.block_size + 1 :] for k in keys]
        self.cached_values = [v[:, -self.block_size + 1 :] for v in values]
        last_queries = [q[:, [-1]] for q in queries]
        attn_maps = self.get_attn_maps(last_queries, keys)
        weighted_values = torch.cat(
            [self.attn_dropout(a) @ v for a, v in zip(attn_maps, values)], dim=-1
        )
        z = self.O(weighted_values)
        z = self.out_dropout(z)
        return z
```

20 Basic Memory, FLOPs, Communication, and Latency

The essentials of inference-time math, much of it based on [40].

Naive Inference Processing a single `(B, S, D)`-shaped tensor to generate a single next input costs the $2BSN_{\text{params}}$ FLOPs we found for the forwards-pass in Sec. 7 (assuming $S \lesssim D$). Memory costs just come from the parameters themselves: $M_{\text{infer}}^{\text{naive}} = pN_{\text{params}}$. Per the analysis of App. D, naive inference is generally compute-bound and so the per-token-latency is approximately[82] $2BSN_{\text{params}}/\lambda_{\text{FLOP/s}}$ where the FLOPs bandwidth in the denominator is again defined in App. D.

kv-Cache Inference The FLOPs requirements for the hidden-dimension matrix multiplies during generation are $2BN_{\text{params}}$, since we are only processing a single token, per previous results. This is in addition to the up-front cost of $2BSN_{\text{params}}$ for the prefill. But, the memory requirements are raised to

$$M_{\text{infer}}^{\text{kv-cache}} = pN_{\text{params}} + 2pBDLS/T \ . \tag{20.1}$$

Inference now has a computational-intensity of

$$\frac{C_{\text{infer}}^{\text{kv-cache}}}{M_{\text{infer}}^{\text{kv-cache}}} \sim \frac{BD}{S} \ , \tag{20.2}$$

dropping $\mathcal{O}(1)$ factors, is now memory-bound (again, see App. D), and has per-token-latency of approximately $M_{\text{infer}}/\lambda_{\text{mem}}$, unless the batch-size is very large.

[82] Assuming we do the naive thing here and generate the next token in a similarly naive way, shifting over the context window.

Intra-Node Communication For T-way tensor parallelism, two `AllReduce`s are needed, one for each `MLP` and each `CausalAttention` layer, where each accelerator is sending $pBDS$ bytes of data (see Sec. 10.3). This requires a total of $4\,(T-1)\,pBDS/T \approx 4pBDS$ bytes to be transferred between workers in the tensor-parallel group (see Foot. 91), taking a total of $\sim 4pBDLS/\lambda_{\text{comms}}$ time for the model as a whole. For an A100 80GiB, `torch.float16` setup, this is $\sim BDS \times 10^{-11}$ sec

Latency TODO

21 Case Study: Falcon-40B

Let's work through the details of the kv-cache for Falcon-40B[83] with $D = 8192$, $L = 60$, $S = 2048$. In half, $p = 2$ precision, the model weights just about fit on an 80GiB A100, but this leaves no room for the cache, so we parallelize T ways across T GPUs, assumed to be on the same node. The total memory costs are then

$$M_{\text{total}} \approx \frac{80\text{GiB} + 4\text{GiB} \times B}{T} \,. \tag{21.1}$$

This means that in order to hit the compute-bound threshold of $B \sim 200$ (see App. D) we need at least $T = 4$ way parallelism. Taking $T = 4$, and running at capacity with $B \sim 200$ so that we are compute-bound, the per-iteration latency from computation alone is approximately $\frac{2BN_{\text{params}}}{\lambda_{\text{FLOP/s}}T}$ ~13ms, i.e. we can give ~200 customers about ~75 tokens-per-second at this rate[84], if this were the only latency consideration.

[83] Falcon actually uses multi-query attention, which changes the computations here, but we will pretend it does not in this section for simplicity.

[84] Average human reading speed is about $\sim$ 185 words/minute, or $\sim$ 4tokens/sec.

A Conventions and Notation

We loosely follow the conventions of [5]. Common parameters:

- A: number of attention heads
- B: microbatch size
- C: compute (FLOPs)
- D: the hidden dimension size
- E: expansion factor for MLP layer (usually $E = 4$)
- H: D/A, the head dimension size
- K: the block size (maximum sequence length[85])
- L: number of transformer layers
- N_{params}: total number of model parameters
- N_{ex}: number of experts for MoE models.
- P: pipeline parallel size
- S: input sequence length
- T: tensor parallel size
- V: vocabulary size
- t: various timescales
- p: the precision of the elements of a tensor in bytes
- λ: various rates, e.g. λ_{mem} is memory bandwidth

Where it makes sense, we try to use the lower-case versions of these characters to denote the corresponding indices on various tensors. For instance, an input tensor with the above batch size, sequence length, and vocabulary size would be written as x_{bsv}, with $b \in \{0, \ldots, B-1\}$, $s \in \{0, \ldots, S-1\}$, and $v \in \{0, \ldots, V-1\}$ in math notation, or as `x[b, s, v]` in code.

Typical transformers belong to the regime

$$V \gg D, S \gg L, A \gg P, T \ . \tag{A.1}$$

For instance, GPT-2 and GPT-3 [2, 3] have $V \sim \mathcal{O}\left(10^4\right)$, $S, L \sim \mathcal{O}\left(10^3\right)$, $L, A \sim \mathcal{O}\left(10^2\right)$. We will often assume also assume that[86] $S \lesssim D$ or the weaker[87] $BS \lesssim D$.

[85] In the absence of methods such as ALiBi [41] can be used to extend the sequence length at inference time.

[86] This condition ensures that the $\mathcal{O}\left(S^2\right)$ FLOPs cost from self-attention is negligible compared to $\mathcal{O}\left(D^2\right)$ contributions from other matrix multiplies. It should be noted that in Summer 2023 we are steadily pushing into the regime where this condition does *not* hold.

[87] This condition ensures that the cost of reading the $\mathcal{O}\left(D^2\right)$ weights is more than the cost of reading in the $\mathcal{O}\left(BSD\right)$ entries of the intermediate representations.

As indicated above, we use zero-indexing. We also use `python` code throughout[88] and write all ML code using standard `torch` syntax. To avoid needing to come up with new symbols in math expressions we will often use expressions like $x \leftarrow f(x)$ to refer to performing a computation on some argument (x) and assigning the result right back to the variable x again.

Physicists often joke (half-seriously) that Einstein's greatest contribution to physics was his summation notation in which index-sums are implied by the presence of repeated indices and summation symbols are entirely omitted. For instance, the dot product between two vectors would be written as

$$\vec{x} \cdot \vec{y} = \sum_i x_i y_i \equiv x_i y_i \tag{A.2}$$

We use similar notation which is further adapted to the common element-wise deep-learning operations. The general rule is that if a repeated index appears on one side of an equation, but not the other, then a sum is implied, but if the same index appears on both sides, then it's an element-wise operation. The Hadamard-product between two matrices A and B is just

$$C_{ij} = A_{ij} B_{ij} \ . \tag{A.3}$$

Einstein notation also has implementations available for `torch`: see this blog post on `einsum` or the `einops` package. We strive to write all learnable weights in upper case.

In particular, we use `einops` notation for concatenation and splitting: $A_c = A_{(de)} = B_{de}$[89]. We will sometimes use a bar to indicate tensors which are derived from other tensors through such splitting operations, usually in the context of tensor-sharding where devices only locally hold some shard of the tensor. In this context, only some of the dimensions will be sharded across devices, and we may also put a bar over the corresponding sharded index. For instance, consider a two-dimensional tensor M_{ab} of shape `M.shape=(A, B)`: sharding this tensor across two devices across the final index results in a tensor $\bar{M}_{a\bar{b}}$ which is of shape `M_bar.shape=(A, B/2)` on each device. As here, we will sometimes use bars to denote indices which are sharded over different devices.

We also put explicit indices on operators such as `Softmax` to help clarify the relevant dimension, e.g. we would write the softmax operation over the b-index of some batched tensor $x_{bvd\ldots}$ as

$$s_{bvd\ldots} = \frac{e^{x_{bvd\ldots}}}{\sum_{v=0}^{v=V-1} e^{x_{bvd\ldots}}} \equiv \texttt{Softmax}_v \ x_{bvd\ldots} \ , \tag{A.4}$$

indicating that the sum over the singled-out v-index is gives unity.

B Collective Communications

A quick refresher on common distributed communication primitives. Consider R ranks with tensor data $x^{(r)}$ of some arbitrary shape `x.shape`, which takes up M bytes of memory, where r labels the worker and any indices on the data are suppressed. For collectives which perform an operation

[88] Written in a style conducive to latex, e.g. no type-hints and clarity prioritized over optimization.

[89] The indexing is all row-major: if A_i is I-dimensional, $i \in \{0, \ldots, I-1\}$, then if we split this index as $A_i = A_{(jk)} \equiv \bar{A}_{jk}$, then the indices j, k will range over $j \in \{0, \ldots, J\}$, $k \in \{0, \ldots, K\}$ with $I = J \times K$ and where numerically $i = j \times K + k$. More complex cases follow by induction.

over a specific dimension, the `torch` convention is that it operates over `dim=0`. The $r = 0$ worker is arbitrarily denoted the *chief*. Some operations are easiest to describe by forming the logical super-tensor `X=torch.stack([x0, x1, ...], dim=0` of shape `X.shape==Size(R, ...)` such that the tensor on rank `r` is `x=X[r]`. Then, the primitive operations are:

- `Broadcast`: all workers receive the chief's data, $x^{(0)}$.
- `Gather`: all workers communicate their data x_n to the chief, e.g. in a concatenated array $[x^0, x^1, \ldots, x^{R-1}]$. E.g., the chief gets `x_out = X.reshape(R*X.shape[1], X.shape[2:])`.
- `Reduce`: data is `Gather`-ed to the chief, which then performs some operation (`sum`, `max`, `concatenate`, etc.) producing a new tensor x' on the chief worker. E.g., for `sum` the chief gets `x_out = X.sum(dim=0)`.
- `ReduceScatter`: a reducing operation (e.g. `sum`) is applied to the $x^{(r)}$ to produce a x' of the same shape (e.g. $x' = \sum x^{(r)}$) and each worker only receives a $1/R$ slice (and hence M/R byte) of the result[90]. A ring implementation sends $M \times \frac{R-1}{R}$ bytes over each link in the ring. E.g., for `sum` rank `r` gets output `x_out = X.sum(dim=0).tensor_split(R, dim=0)[r]`.
- `AllGather`: all data $x^{(r)}$ is communicated to all workers; each worker ends up with the array $[x^0, x^1, \ldots, x^{R-1}]$. Functionally equivalent to a `Gather` followed by `Broadcast`. A ring implementation sends $M \times (R-1)$ bytes over each link in the ring. E.g., all ranks get `x_out = X.reshape(R*X.shape[1], X.shape[2:])`.
- `AllReduce`: all workers receive the same tensor x' produced by operating on the $x^{(r)}$ with `sum`, `mean`, etc. Functionally equivalent to a `Reduce` followed by `Broadcast`, or a `ReduceScatter` followed by a `AllGather` (the more efficient choice[91]). In the latter case, the total cost is $2M \times \frac{R-1}{R}$, due to `AllReduce`-ing the initial M-sized data, and then `AllGather`-ing the M/R-sized reductions. E.g., for `sum` all ranks get `x_out = X.sum(dim=0)`.
- `Scatter`: One worker gives shards of a tensor to all workers. If the worker is scattering tensor T_x over the given index, a `Scatter` effectively shards this as $T_x \longrightarrow T_{(\bar{r}y)}$, each worker getting a $\bar{r}$-shard. If `x` is the chief's data, rank `r` receives `x_out = x.tensor_split(R, dim=0)[r]`.
- `AllToAll`: All workers receive shards of all others worker's tensors. If every worker has a tensor $T_{\bar{r}y}$, for one value of $\bar{r}$, which we imagine came from a sharding a tensor $T_x = T_{(\bar{r}y)}$, then an `AllToAll` over the y index produces produces the tensor $T_{z\bar{r}}$ defined by $T_{z\bar{r}} = T_x$ on all workers. E.g. rank `r` receives `x_out = X.reshape(X.shape[1], R, X.shape[:2])[:,r]`.

[90] Note that `AllGather` and `ReduceScatter` are morally conjugate to each other. In the former, each worker ends up with R times as much data as they started with, while in `ReduceScatter` they end up with $1/R$ of their initial data. One is nearly a time-reversed version of the other, which is a way of remembering that they have the came communication cost. They also compose to produce an output of the same initial size, as in `AllReduce`.

[91] The former strategy scales linearly with the number of worker, while the latter strategy underlies "ring" `AllReduce` which is (nearly) independent of the number of workers: if each worker carries data of size D which is to be `AllReduce`-d, a total of $\frac{2(R-1)D}{R}$ elements need to be passed around. See this blog post for a nice visualization or [42] for a relevant paper.

C Hardware

Basic information about relevant hardware considerations. Much of the following is from the NVIDIA docs.

C.1 NVIDIA GPU Architecture

NVIDIA GPUs consist of some amount of relatively-slow off-chip DRAM memory[92], relatively-fast on-chip SRAM, and a number of **streaming multiprocessors** (SMs) which perform the parallel computations. Inside more-recent GPUs, the SMs carry both "CUDA cores" and "Tensor cores", where the latter are used for matrix-mulitiplies and the former for everything else.

A few numbers of primary importance:

- The rate at which data can be transferred from DRAM to SRAM (λ_{mem})
- The number of FLOP/s, which is more fundamentally computed by multiplying the number of SMs by the FLOPS/cycle of each SM for the specific operation under consideration (see the NVIDIA docs) by the clock rate: $N_{\text{SM}} \cdot \lambda_{\text{FLOPs/cycle}} \cdot \lambda_{\text{clock}}$

The terminology and structure of the memory hierarchy is also important to understand. Types of memory, from slowest to fastest:

- **Global** memory is the slow, but plentiful, off-chip DRAM. It is the type of memory typically used as kernel arguments
- **Constant** memory is read only and accessible by all threads in a given block. The size of arrays in constant memory must be known at compile time
- **Local Memory** is similarly slow to global memory, but more plentiful than register memory, and privately to individual threads and is allocated from within a kernel. When registers run out, local memory fills the gap
- **Shared** memory is shared between all threads in a given block. Shared memory is effectively a user-controlled cache. The size of arrays in shared memory must be known at compile time
- **Registers** hold scalar values and small tensors whose values are known at compile time. They are local to each thread and they are plentiful since each thread needs its own set of registers: $65{,}536 = 2^{16}$ registers per SM an A100.

An excellent video overview of CUDA and NVIDIA GPU architecture which covers some of the above is here.

C.2 CUDA Programming Model

The CUDA programming model uses a hierarchy of concepts:

[92] This is the number usually reported when discussing a given GPU, e.g. 32GiB for the top-of-the-line A100

- **Threads** are the fundamental unit of execution[93] which each run the same CUDA **Kernel**, or function, on different data inputs in parallel. Threads within the same block (below) may share resources, like memory, and may communicate with each other. Individual threads are indexed through the **threadIdx** variable, which has `threadIdx.{x, y, z}` attributes with `threadIdx.x` in `0, ..., blockDim.x -1` and similar.
- Threads (and hence warps) are organized into 3D **blocks**. The size and indices of the blocks can be accessed through the `blockDim` and `blockIdx` variables, respectively, with `blockIdx.x` in `0, ..., gridDim.x - 1`. `blockDim.x *blockDim.y * blockDim.z` total threads run in a block.
- Blocks are organized into 3D **groups**. The size of the gird dimensions can be accessed through the `gridDim` variable, with similar attributes to the above.
  `gridDim.x * gridDim.y * gridDim.z` total blocks run in a grid.

The number of threads which can be launched in a given block is hardware limited; A100 80GiB GPUs can run up to 1024 threads in a SM at a time (32 blocks with 32 threads each), for instance. Hence, block and grid sizes need to be adjusted to match the problem size. There are also important memory access considerations here. The 1024 threads which can be launched can also read sequentially from memory and efficient usage implies that choosing the block size such that we are doing these reads as often as possible is ideal.

C.3 NVIDIA GPU Stats

Summary of some relevant NVIDIA GPU statistics:

GPU	Memory	$\lambda_{\text{FLOP/s}}$	λ_{mem}	λ_{math}	λ_{comms}
A100	80GiB	312 TFLOP/s	2.0 TiB/s	156 FLOPS/B	300 GiB/s
A100	40GiB	312 TFLOP/s	1.6 TiB/s	195 FLOPS/B	300 GiB/s
V100	32GiB	130 TFLOP/s	1.1 TiB/s	118 FLOPS/B	16 GiB/s

where

- $\lambda_{\text{FLOP/s}}$ is flops bandwidth (for `float16`/`bfloat16` multiply-accumulate ops)
- λ_{mem} is memory bandwidth
- $\lambda_{\text{math}} = \frac{\lambda_{\text{FLOP/s}}}{\lambda_{\text{mem}}}$ is math bandwidth
- λ_{comms} is one-way communication bandwidth

A useful approximate conversion rate is that 1 TFLOP/s $\approx$ 100 PFLOP/day.

Important practical note: the $\lambda_{\text{FLOP/s}}$ numbers should be taken as aspirational. Out-of-the box, `torch.float16` matrix-multiplies in `torch` with well-chosen dimensions tops out around $\sim$ 250 FLOPS/s

[93] Threads are always physically launched in **Warps** which consist of 32 threads.

D Compute-bound vs Memory-bound

If your matrix-multiplies are not sufficiently large on, you are wasting resources [43]. The relevant parameters which determine sufficiency are $\lambda_{\text{FLOP/s}}$ and λ_{mem}, the FLOPs and memory bandwidth, respectively. The ratio $\lambda_{\text{math}} \equiv \frac{\lambda_{\text{FLOP/s}}}{\lambda_{\text{mem}}}$ determines how many FLOPS you must perform for each byte loaded from memory; see App. C.3. If your computations have a FLOPs/B ratio which is larger than λ_{math}, then you are compute-bound (which is good, as you're maximizing compute), and otherwise you are memory(-bandwidth)-bound (which is bad, since your compute capabilities are idling). The FLOPs/B ratio of your computation is sometimes called the **compute intensity** or **arithmetic intensity**. When compute bound, a process takes time $\sim F/\lambda_{\text{FLOP/s}}$, while memory-bound processes take time[94] $\sim M/\lambda_{\text{mem}}$.

D.1 Matrix-Multiplications vs. Element-wise Operations

For instance, to multiply a `(B, S, D)`-shaped tensor z_{bsd} by a `(D, D)`-shaped weight-matrix $W_{dd'}$, $p\left(BDS + D^2\right)$ bytes must be transferred from DRAM to SRAM at a rate λ_{mem}, after which we perform $2BSD^2$ FLOPs, and write the `(B, S, D)`-shaped result back to DRAM again, for a ratio of

$$\frac{1}{p}\frac{BDS}{2BS + D} \text{ (FLOPs/B) } . \tag{D.1}$$

We want to compare this against λ_{math}, which from App. C.3 we take to be $\mathcal{O}\left(100 \text{ FLOPs/B}\right)$, and plugging in any realistic numbers, shows that such matrix-multiplies are essentially always compute-bound. Compare this to the case of some element-wise operation applied to the same z_{bsd} tensor whose FLOPs requirements are $\sim C \times BDS$ for some constant-factor $C \ll S, D$. Then, then FLOPS-to-bytes ratio is $\sim \frac{C}{p}$, which is *always* memory-bound for realistic values of C. The moral is to try and maximize the number of matrix-multiplies and remove as many element-wise operations that you can get away with.

D.2 Training vs. Inference

Finally, we note that the above has implications for the Transformers architecture as a whole, and in particular it highlights the difficulties in efficient inference. Under the assumptions of Sec. 7, $\sim \mathcal{O}\left(BSN_{\text{params}}\right)$ total FLOPs needed during training, while the number of bytes loaded from and written to memory are $\mathcal{O}\left(BDLS + N_{\text{params}}\right) \sim \mathcal{O}\left(\frac{BSN_{\text{params}}}{D} + N_{\text{params}}\right)$ which is $\mathcal{O}\left(N_{\text{params}}\right)$ for not-super-long sequence lengths. The arithmetic intensity is therefore $\mathcal{O}\left(BS\right)$ and so training is compute-bound in any usual scenario, even at small $B \sim \mathcal{O}\left(1\right)$ batch sizes (as long as individual operations in the network don't suffer from outlandish memory-boundedness). The problem during inference is that (if using the kv-cache; see Sec. 19) we only need to process a *single* token at a time and so $S \longrightarrow 1$ in the numerator in the preceding, while the denominator is also weighed down by the kv-cache in the attention layers.

[94] Note that the time is not additive, e.g. compute-bound tasks do not take time $\sim F/\lambda_{\text{FLOP/s}} + M/\lambda_{\text{mem}}$ because they are not sequential: compute and memory-communications can be concurrent.

In more detail, the `MLP` layers just process $S = 1$ length tensors during generation, but are insensitive to the kv-cache, so their intensity comes from just setting $S = 1$ in the above,

$$\sim \frac{BD}{B+D} \ , \tag{D.2}$$

dropping $\mathcal{O}(1)$ factors now, while the attention layers have a ratio of the form

$$\sim \frac{BDS + BD^2}{BD + D^2 + BDS} \ , \tag{D.3}$$

where the last term in the denominator is due to the cache. Now at small $B \sim \mathcal{O}(1)$ batch sizes, both intensities reduce to $\mathcal{O}(B)$, which is insufficient to be compute-bound. In the large $B \gtrsim D/S$ limit, they at least become $\mathcal{O}(D)$ and $\mathcal{O}\left(1 + \frac{D}{S}\right)$, respectively, which may be enough to be compute-bound, but it's hard to even get into this regime. Note, the importance of the ratio D/S. The hidden dimension fixes the context length scale at which inference can never be compute-bound, in the absence of additional tricks not considered here[95].

D.3 Intra- and Inter-Node Communication

For intra-node communication, GPUs are connected by either PCIe or NVLink, generally.

- NVLink interconnects are continually updated and achieve speeds of $\lambda_{\text{comm}}^{\text{intra}} \sim 300$ GiB/s.

For inter-node communication, nodes are often connected by:

- InfiniBand apparently also achieves speeds $\lambda_{\text{comm}}^{\text{intra}} \sim 100$ GiB/s? Haven't found a clear reference. But in any case, the bandwidth is divided amongst the GPUs in the node, leading to a reduction by ~ 8.

E Batch Size, Compute, and Training Time

The amount of compute directly determines the training time, but not all ways of spending compute are equivalent. We follow the discussion in [44] which gives a rule of thumb for determining the optimal batch size which is sometimes used in practice. The basic point is that all of the optimization steps take the gradient $\mathbf{g}$ as an input, and since the gradient is the average over randomly selected datapoints, steps are more precise as the batch size increases (with diminishing returns, past a certain point, but the computational cost also rises with batch size, and a balance between the two concerns should be struck.

Consider vanilla SGD and study how the training loss changes with each step. We randomly sample B datapoints $x \in \mathcal{D}$ from the dataset through some i.i.d. process[96]. Each corresponding

[95] One such trick: the multi-query attention of Sec. 2.2 improves everything a factor of A: the large batch regime is $B \gtrsim \frac{D}{AS}$ and the intensity ratio becomes $\mathcal{O}\left(1 + \frac{D}{AS}\right)$. An analysis equivalent to the one performed here can be found in the original paper [7].

[96] The below uses sampling with replacement, while in practice we sample without replacement, but the different is negligible for all practical cases.

gradient $\mathbf{g}(x) = \partial_w \mathcal{L}(w, x)$ is itself a random variable whose average is the true gradient across the entire dataset $\bar{\mathbf{g}}$ and we take the variance to be

$$\mathrm{Var}[\mathbf{g}(x), \mathbf{g}(x')] = \Sigma \tag{E.1}$$

for some matrix Σ with (supressed) indices spanning the space of model weights. Taking instead the mean of a sum of such estimates, $\mathbf{g}_B \equiv \frac{1}{B}\sum_{x\in\mathcal{B}} \mathbf{g}(x)$, the mean stays the same, but the variance reduces in the usual way: $\mathrm{Var}[\mathbf{g}_B(x), \mathbf{g}_B(x')] = \Sigma/B$.

Study the mean loss across the entire dataset: $\mathcal{L}(w) = \langle \mathcal{L}(w, x)\rangle$. Using SGD we take a step $w \longrightarrow w - \eta \mathbf{g}_B$ and change the loss as

$$\mathcal{L}(w - \eta \mathbf{g}_B) = \mathcal{L}(w) - \eta \bar{\mathbf{g}} \cdot \mathbf{g}_B + \frac{1}{2}\mathbf{g}_B \cdot H \cdot \mathbf{g}_B + \mathcal{O}\left(\mathbf{g}_B^3\right) \ , \tag{E.2}$$

where H is the true hessian of the loss over the entire dataset at this value of the weights. Taking the expectation value and minimizing the results over η gives the optimal choice:

$$\eta_\star = \frac{\eta_{\max}}{1 + \frac{B_{\mathrm{noise}}}{B}} \ , \quad \eta_{\max} \equiv \frac{\bar{\mathbf{g}}^2}{\bar{\mathbf{g}} \cdot H \cdot \bar{\mathbf{g}}}, \quad B_{\mathrm{noise}} \equiv \frac{\mathrm{Tr}\, H \cdot \Sigma}{\bar{\mathbf{g}} \cdot H \cdot \bar{\mathbf{g}}} \ . \tag{E.3}$$

Notably, the above supports the usual rule of thumb that the learning rate should be increased proportionally to the batch size, at least whenever $B \ll B_{\mathrm{noise}}$. The diminishing returns of pushing batch sizes past B_{noise} are also evident. In practice it is too expensive to compute the Hessian, but thankfully the entirely unjustified approximation in which the Hessian is multiple of the identity such that

$$B_{\mathrm{noise}} \approx B_{\mathrm{simple}} \equiv \frac{\mathrm{Tr}\,\Sigma}{\bar{\mathbf{g}}^2} \ , \tag{E.4}$$

is somehow a decent approximation empirically, and an estimator can be created for B_{noise} in a data-parallel setup; see [44] or Katherine Crowson's implementation or neox for more.

We can further characterize the trade-off between compute and optimization steps. The expected decrease in loss per update is then

$$\langle \delta \mathcal{L} \rangle \approx \frac{\eta_{\max}}{1 + \frac{B_{\mathrm{noise}}}{B}} \bar{\mathbf{g}}^2 + \mathcal{O}\left(\eta_{\max}^2\right) \ , \tag{E.5}$$

that is, we would need $1 + \frac{B_{\mathrm{noise}}}{B}$ times as many SGD steps to make the same progress we would have as compared to full-batch SGD. If $S_{\min}$ is the number of steps that would have been needed for full-batch SGD, we would need $S = S_{\min} + S_{\min}\frac{B_{\mathrm{noise}}}{B}$ steps for minibatch SGD. The total number of examples seen is correspondingly $E = S_{\min} \times (B_{\mathrm{noise}} + B) \equiv E_{\min} + S_{\min} B$, and so we see the trade-off between SGD steps S and compute E alluded to above. These relations can be written as[97]

$$\left(\frac{S}{S_{\min}} - 1\right)\left(\frac{E}{E_{\min}} - 1\right) = 1 \tag{E.6}$$

[97] The analysis here is simplified in that it assumes that the noise scale and the chosen batch size are both time-independent. There is confusing logic treating the more general case where both B_{noise} and B vary with step in [44], but in any case, the ultimate relations they use are effectively the same.

which represent hyperbolic Pareto frontier curves. So, solutions are of the form $S = (\alpha + 1) S_{\min}$, $E = \left(\frac{1}{\alpha} + 1\right) E_{\min}$ and since $E = BS$ the corresponding batch size is $B_{\text{crit}} \equiv \frac{1}{\alpha} B_{\text{noise}}$. The parameter α characterizes how much you value the trade-off between these two factors and a reasonable balance is the $\alpha = 1$ solution for which $S = 2S_{\min}$, $E = 2E_{\min}$ and $B_{\text{crit}} = B_{\text{noise}}$ exactly.

Correspondingly, in [44] they suggest training at precisely this batch size. But it seems much more relevant to balance time against compute directly, rather than optimization steps vs compute. Modeling the total training time by $T \approx S\left(\kappa B + \sigma\right)$ for some κ, σ to model compute costs[98], then the above is equivalent to

$$T = \frac{\left(E_{\min} + S_{\min} B\right)\left(\kappa B + \sigma\right)}{B} \ . \tag{E.7}$$

which has a minimum at

$$B = \sqrt{\frac{\sigma E_{\min}}{\kappa S_{\min}}} \ . \tag{E.8}$$

for which the total time is

$$T_{\min} = \left(\sqrt{\kappa E_{\min}} - \sqrt{\sigma S_{\min}}\right)^2 \ . \tag{E.9}$$

In comparison, the total time for the $B_{\text{crit}} = \frac{E_{\min}}{S_{\min}}$ strategy of [44] gives $T_{\min} = 2\left(\kappa E_{\min} + \sigma S_{\min}\right)$ which is a factor of $\frac{2}{1 - \frac{\sqrt{\sigma \kappa B_{\text{noise}}}}{\kappa B_{\text{noise}} + \sigma}}$ larger. So, this seems like a better choice of optimal batch size, if you value your time.

F Initialization, Learning Rates, μ-Transfer etc

A quick review of common initialization strategies and arguments for learning rate choices and μ-transfer. We follow some mix of [8, 45–47].

The core principles are that, at least at the early stages of training, we attempt to make identified activations in different blocks have approximately equal statistics[99] and demand that for each training step the contribution of each weight's change to the architecture's outputs should be roughly equal for identified weights in different blocks. Further, this should occur for all choices of architectural parameters. In particular large-width, $D \longrightarrow \infty$ limit should be D-independent at first non-trivial order, which is the easiest limit to reason about.

We mostly specialize to very simple cases in the following: MLP-only models which may have trivial non-linearities.

F.1 Wide Models are Nearly Gaussian

First we discuss the justification of an assumption we make throughout: the outputs every block (suitably defined) at initialization is approximately normally distributed.

[98] Computation and communication costs each scale with B, the optimizer step does not (and maybe some overhead?), for instance.

[99] Assuming there is some regular block structure with a corresponding natural identification between weights and activations in different blocks.

Take our model to be $z_i^\ell = W_{ij}^\ell \phi\left(z_j^{\ell-1}\right)$ where the inputs z_i^0 are i.i.d. Gaussian-normally distributed[100]: $\mathbf{E}\left[z_i^0\right] = 0$, $\mathbf{E}\left[z_i^0 z_j^0\right] = \delta_{ij}$. Here, $i \in \{0, \ldots, D-1\}$ and the batch and any other indices are suppressed.

Examine the statistics of the first layer. Choosing the weights to be normally distributed as well, with $\mathbf{E}\left[W_{ij}^\ell\right] = 0$, $\mathbf{E}\left[W_{ij}^\ell W_{jk}^\ell\right] = \frac{C_\ell}{D}\delta_{ij}$ for some C_ℓ it straightforward to show that

$$\begin{aligned} \mathbf{E}\left[z_i^1\right] &= 0 \\ \mathbf{E}\left[z_i^1 z_j^1\right] &= C_1 \delta_{ij} \langle \phi(z)^2 \rangle \end{aligned} \tag{F.1}$$

where $\langle \phi(z)^n \rangle \equiv \int \mathrm{d}\, \rho(z) \phi(z)^n$ with $\rho(z)$ a single-variable standard normal Gaussian[101] (the D in the denominator was chosen to counteract a factor of D from an index sum), which are all some $\mathcal{O}(1)$, D-independent numbers.

The first two moments can therefore be made Gaussian-normal-like by choosing $C_1 = 1/\mathbf{E}\left[\phi(z)\phi(z)\right]$. Since this can always be done, the first non-trivial test of non-Gaussianity is the four-point function (the three-point function vanishes by symmetries). The connected four-point function[102] show the presence of non-gaussianity most directly. Symmetries fix the result to be of the form

$$\mathbf{E}\left[z_i^\ell z_j^\ell z_k^\ell z_l^\ell\right]_c = V_4^\ell \delta_{ij}\delta_{kl} + \text{perms}\ , \tag{F.2}$$

for some coefficient V_4^ℓ for all ℓ. We can fix the coefficient by computing the term, say, where $i = j, k = l, i \neq k$. The result for the $\ell = 1$ layer is:

$$V_4^{\ell=1} = \frac{C_1^2}{D^2}\left[\mathbf{E}\left[\left(\phi(z^0)\right) \cdot \phi(z^0)\right)^2\right] - \left(\mathbf{E}\left[\phi(z^0)) \cdot \phi(z^0)\right)\right]\right)^2\right] \tag{F.3}$$

where the expectation is over the distribution of z_i^0 and W_{ij}^1 and the dot-product is over hidden-dimension indices. This can be written in terms of the single-variable expectation values $\langle \phi(z)^n \rangle$ with the result:

$$V_4^{\ell=1} = \frac{C_1^2}{D}\left(\langle \phi(z)^4 \rangle - \langle \phi(z)^2 \rangle^2\right)\ . \tag{F.4}$$

So, there is indeed non-gaussianity (even for a linear network $\phi(x) = x$) and is it of $\mathcal{O}\left(\frac{1}{D}\right) \ll 1$. Perhaps unsurprisingly, we can continue on to deeper layers via perturbation theory and find $V_4^\ell \sim \mathcal{O}\left(\frac{\ell}{D}\right)$; the non-linearity is additive in the $L/D \ll 1$ regime. Similar results also hold for higher-order, even-point functions.

We will assume that arguments like this can be generalized for all networks under consideration: many activations are approximately Gaussian-normally distributed, after appropriately tuning initialization scales. Demonstrating this rigorously is a central goal of the Tensor Programs work [8].

[100] It may be that these inputs come from some transformation of other data elements. For example, for an LLM the z_i^0 come from looking up the normally-distributed embedding vectors corresponding to the relevant tokens in the sequence.

[101] This is similar notation as used in [45], with $\mathbf{E}[\cdot]$ being a multivariate expectation value and $\langle\cdot\rangle$ an expectation value over a 1D distribution.

[102] Also known as the cumulant: $\mathbf{E}\left[z_i^\ell z_j^\ell z_k^\ell z_l^\ell\right]_c \equiv \mathbf{E}\left[z_i^\ell z_j^\ell z_k^\ell z_l^\ell\right] - \mathbf{E}\left[z_i^\ell z_j^\ell\right]\mathbf{E}\left[z_k^\ell z_l^\ell\right] - \text{perms}$.

G Cheat Sheet

Collecting all of the most fundamental equations, given to various degrees of accuracy.

Number of model parameters:

$$N_{\text{params}} = (4+2E)LD^2 + VD + \mathcal{O}(DL) \approx (4+2E)\,LD^2\ , \tag{G.1}$$

assuming no sharding of the embedding matrix.

Training Memory costs for mixed-precision training:

$$\begin{aligned} M_{\text{model}} &= p_{\text{model}} N_{\text{params}} \\ M_{\text{optim}} &= (s_{\text{states}}+1) \times p_{\text{master}} N_{\text{params}} \\ M_{\text{act}}^{\text{total}} &= \frac{2BDLS\,(p(E+4)+1)}{T} + \frac{ABLS^2\,(2p+1)}{T} + \mathcal{O}(BSV) \end{aligned} \tag{G.2}$$

where s_{states} is the number of optimizer states, e.g. $s=0$ for SGD and $s=2$ for Adam. FLOPs total:

$$F_{\text{total}}^{\text{model}} \approx 12BDLS\,(S+(2+E)\,D)\,. \tag{G.3}$$

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," arXiv:1706.03762 [cs.CL]. 5, 6

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog* **1** (2019) no. 8, 9. 5, 59

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," arXiv:2005.14165 [cs.CL]. 59

[4] OpenAI, "Gpt-4 technical report," arXiv:2303.08774 [cs.CL]. 5

[5] V. Korthikanti, J. Casper, S. Lym, L. McAfee, M. Andersch, M. Shoeybi, and B. Catanzaro, "Reducing activation recomputation in large transformer models," arXiv:2205.05198 [cs.LG]. 5, 28, 32, 44, 59

[6] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu, "On layer normalization in the transformer architecture," arXiv:2002.04745 [cs.LG]. 6

[7] N. Shazeer, "Fast transformer decoding: One write-head is all you need," arXiv:1911.02150 [cs.NE]. 7, 14, 65

[8] G. Yang, E. J. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, and J. Gao, "Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer," arXiv:2203.03466 [cs.LG]. 8, 67, 68

[9] N. Shazeer, "Glu variants improve transformer," arXiv:2002.05202 [cs.LG]. https://arxiv.org/abs/2002.05202. 14

[10] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, "Gqa: Training generalized multi-query transformer models from multi-head checkpoints," arXiv:2305.13245 [cs.CL]. 14

[11] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," arXiv:2307.09288 [cs.CL]. 14

[12] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," arXiv:2204.02311 [cs.CL]. 14

[13] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," arXiv:2104.09864 [cs.CL]. 15

[14] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "Flashattention: Fast and memory-efficient exact attention with io-awareness," arXiv:2205.14135 [cs.LG]. 16, 26

[15] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," arXiv:2307.08691 [cs.LG]. 16

[16] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," arXiv:2006.16236 [cs.LG]. https://arxiv.org/abs/2006.16236. 19

[17] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *CoRR* **abs/2111.00396** (2021) , 2111.00396. https://arxiv.org/abs/2111.00396. 20

[18] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," arXiv:2312.00752 [cs.LG]. https://arxiv.org/abs/2312.00752. 20, 21

[19] T. Dao and A. Gu, "Transformers are ssms: Generalized models and efficient algorithms through structured state space duality," arXiv:2405.21060 [cs.LG]. https://arxiv.org/abs/2405.21060. 24

[20] G. E. Blelloch, "Prefix sums and their applications,". https://www.cs.cmu.edu/~guyb/papers/Ble93.pdf. 25

[21] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," arXiv:1710.03740 [cs.AI]. 27

[22] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," arXiv:1606.08415 [cs.LG]. 30

[23] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," arXiv:2001.08361 [cs.LG]. 34, 35, 36

[24] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, "Training compute-optimal large language models," arXiv:2203.15556 [cs.CL]. 35, 36

[25] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," arXiv:2305.18290 [cs.LG]. https://arxiv.org/abs/2305.18290. 37, 38

[26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv:1707.06347 [cs.LG]. https://arxiv.org/abs/1707.06347. 37

[27] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela, "Kto: Model alignment as prospect theoretic optimization," arXiv:2402.01306 [cs.LG]. https://arxiv.org/abs/2402.01306. 38

[28] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," arXiv:1909.08053 [cs.CL]. 40, 41, 43, 45, 46

[29] H. Liu, M. Zaharia, and P. Abbeel, "Ring attention with blockwise transformers for near-infinite context," arXiv:2310.01889 [cs.CL]. https://arxiv.org/abs/2310.01889. 45

[30] W. Brandon, A. Nrusimha, K. Qian, Z. Ankner, T. Jin, Z. Song, and J. Ragan-Kelley, "Striped attention: Faster ring attention for causal transformers," arXiv:2311.09431 [cs.LG]. https://arxiv.org/abs/2311.09431. 46, 47

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani,

M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv:2010.11929 [cs.CV]. https://arxiv.org/abs/2010.11929. 49

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," arXiv:2103.00020 [cs.CV]. https://arxiv.org/abs/2103.00020. 49

[33] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," arXiv:2101.03961 [cs.LG]. https://arxiv.org/abs/2101.03961. 51

[34] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," arXiv:1701.06538 [cs.LG]. https://arxiv.org/abs/1701.06538. 51

[35] N. Muennighoff, L. Soldaini, D. Groeneveld, K. Lo, J. Morrison, S. Min, W. Shi, P. Walsh, O. Tafjord, N. Lambert, Y. Gu, S. Arora, A. Bhagia, D. Schwenk, D. Wadden, A. Wettig, B. Hui, T. Dettmers, D. Kiela, A. Farhadi, N. A. Smith, P. W. Koh, A. Singh, and H. Hajishirzi, "Olmoe: Open mixture-of-experts language models," arXiv:2409.02060 [cs.CL]. https://arxiv.org/abs/2409.02060. 51

[36] T. Gale, D. Narayanan, C. Young, and M. Zaharia, "Megablocks: Efficient sparse training with mixture-of-experts," arXiv:2211.15841 [cs.LG]. https://arxiv.org/abs/2211.15841. 52

[37] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," arXiv:1904.09751 [cs.CL]. 54

[38] Y. Leviathan, M. Kalman, and Y. Matias, "Fast inference from transformers via speculative decoding," arXiv:2211.17192 [cs.LG]. https://arxiv.org/abs/2211.17192. 54

[39] R. Pope, S. Douglas, A. Chowdhery, J. Devlin, J. Bradbury, A. Levskaya, J. Heek, K. Xiao, S. Agrawal, and J. Dean, "Efficiently scaling transformer inference," arXiv:2211.05102 [cs.LG]. 55

[40] C. Chen, "Transformer inference arithmetic,". https://kipp.ly/blog/transformer-inference-arithmetic/. 57

[41] O. Press, N. A. Smith, and M. Lewis, "Train short, test long: Attention with linear biases enables input length extrapolation," *CoRR* **abs/2108.12409** (2021) , 2108.12409. https://arxiv.org/abs/2108.12409. 59

[42] P. Patarasuk and X. Yuan, "Bandwidth optimal all-reduce algorithms for clusters of workstations," *Journal of Parallel and Distributed Computing* (2009) . 61

[43] H. He, "Making deep learning go brrrr from first principles,". https://horace.io/brrr_intro.html. 64

[44] S. McCandlish, J. Kaplan, D. Amodei, and O. D. Team, "An empirical model of large-batch training," arXiv:1812.06162 [cs.LG]. 65, 66, 67

[45] D. A. Roberts, S. Yaida, and B. Hanin, "The principles of deep learning theory," *CoRR* **abs/2106.10165** (2021) , 2106.10165. https://arxiv.org/abs/2106.10165. 67, 68

[46] S. Yaida, "Meta-principled family of hyperparameter scaling strategies," arXiv:2210.04909 [cs.LG]. https://arxiv.org/abs/2210.04909.

[47] D. Doshi, T. He, and A. Gromov, "Critical initialization of wide and deep neural networks through partial jacobians: General theory and applications," arXiv:2111.12143 [cs.LG]. https://arxiv.org/abs/2111.12143. 67