GeMMMapReduce
or
Commutative Monoids and Memory Efficient Layers in Neural Networks.

Amaru Cuba Gyllensten

May 2025

Abstract

This paper aims at generalizing recent memory and cache-efficient implementations of common layers in Neural Networks, most notably FlashAttentionDao et al. (2022). The main observation is that these layers follow a common pattern, the same which has been used in MapReduce in the Big Data setting, and Monoidal folds in the functional programming setting. We present monoids which enable streaming and efficient implementations of Attention, CrossEntropy, and Two-Layer MLPs, and identify a property of (some) Commutative Monoids which enable these efficient folds.

The purpose of this paper is to illuminate the connection between MapReduce or Monoidal Folds to efficient Layers in Neural Networks, and propose a primitive for folding over the result of (mapped) matrix multiplications, which we call GeMMMapReduce (General Matrix Multiply Map Reduce).

1 Introduction

1.1 Notation

This paper employs liberal use of type and lambda notation borrowed from functional programming. A short primer of this notation is given below.

Types & Terms $a : A$ denotes that the term a has type A. For example, $a : \mathbb{R}$ denotes that the term a is a real number.

Functions $A \to B$ denotes the type of functions from A to B. Lambda functions are written out as $f = x \mapsto x^2 + 3$, meaning $f(x) = x^2 + 3$.

$\to$ is interpreted in a right-associative manner, meaning $A \to B \to C = A \to (B \to C)$, which is isomorphic to $A \times B \to C$.

Product Types $A \times B$ denotes the product type of A and B. For brevity, terms of product types are occasionally written out as tuples. If $a : A$ and $b : B$, we can write $(a, b) : A \times B$, and vice versa. For a term $x : A \times B$, x_0 is the term corresponding to the left-hand type (A), and x_1 corresponds to the right hand-type. We also used named product types, where $\{a : A, b : B\}$ is the type of pairs of A and B indexed by the names a and b, i.e. for $x : \{a : A, b : B\}$ x_a is the A-part of the product, and x_b is the B-part of the product. If y and z are terms of type A and B, the term $x = \{a : y, b : z\}$, i.e. x such that $x_a = y$ and $x_b = z$.

Exponent Types A type A^N where N is an integer denotes the type of N-tuples of A. In other words, $\mathbb{R}^2$ denotes the type of pairs of reals, or points in a plane. More generally, a term $v : A^B$ can be thought of as an array of As indexed by B (or equivalently, $B \to A$), or — when appropriate — a tensor (in the machine learning sense).

1.2 What are commutative monoids?

A monoid S is a tuple $(T, \circ, \mathbf{id})$, where T is the type, and $\circ : T \to T \to T$ is a binary operation. The binary operation must be associative, i.e. $a \circ (b \circ c) = (a \circ b) \circ c$, and $\mathbf{id}$ must be an identity element, i.e. $\mathbf{id} \circ a = a = a \circ \mathbf{id}$. Commutative monoids are monoids where the binary operation is commutative as well as associative, i.e. $a \circ b = b \circ a$.

There are many instances of commutative monoids in modern deep learning:

Sum Summation over real numbers:

$$\begin{aligned} T &= \mathbb{R} \\ x \circ y &= x + y \\ \mathbf{id} &= 0 \end{aligned}$$

WSum Weighted sums of vectors.

$$\begin{aligned} T &= \{v : \mathbb{R}^D, w : \mathbb{R}_{\geq 0}\} \\ x \circ y &= z \\ \text{where } z_w &= x_w + y_w \\ z_v &= \begin{cases} x_v \frac{x_w}{z_w} + y_v \frac{y_w}{z_w}, & z_w > 0 \\ 0, & \text{otherwise} \end{cases} \\ \mathbf{id} &= \{v : 0, w : 0\} \end{aligned}$$

LogWSum Weighted sums of vectors with weights in logspace[1]:

$$
\begin{aligned}
T &= \{v : \mathbb{R}^D, w : \mathbb{R}\} \\
x \circ y &= z \\
\text{where } z_w &= \ln(\exp(x_w) + \exp(y_w)) \\
z_v &= \begin{cases} x_v \exp(x_w - z_w) + y_v \exp(y_w - z_w), & z_w \neq -\infty \\ 0, & \text{otherwise} \end{cases} \\
\mathbf{id} &= \{v : 0, w : -\infty\}
\end{aligned}
$$

Given a commutative monoid S, we can derive a function $fold$ that takes as input a nonempty sequence of Ts and returns the product over all elements in the sequence: $fold : [T] \to T$.

2 Commutative Monoids and Machine Learning

There are a couple of benefits of using folds over commutative monoids in the context of machine learning:

- There is no need to materialize all values in memory during the computation of a fold, enabling better usage of GPU-memory.
- Partial results can be computed and combined in arbitrary order, reducing the need for synchronization.
- *If* the derivative $\frac{d\ x \circ y}{d\ x}$ can be expressed as a function of $x \circ y$ and x, then the same holds for $\frac{d\ fold(X)}{d\ X_i}$, allowing local computation of partial gradients and further reducing the need for communication, computation, and synchronization.

Theorem 1. *Given a commutative monoid* $(T, \circ, \mathbf{id})$ *whose derivative* $\frac{d\ x \circ y}{d\ x}$ *is a function* (D) *of* $x \circ y$ *and* x*, i.e:*

$$\frac{d\ x \circ y}{d\ x} = D(x \circ y, x) \tag{1}$$

We have that for any sequence $X : [T]$*, the derivative of the fold over* X *and any element* X_i *is the same function* D *of* $fold(X)$ *and* X_i*:*

$$\frac{d\ fold(X)}{d\ X_i} = D(fold(X), X_i)$$

[1] This fold applies to attention: $\text{softmax}(a)V = x_0 \circ x_1 \circ \cdots$, where $x_i = \{v = V_i, w = a_i\}$

Proof. Given a sequence X, we let $X_{\neg i}$ be the sequence X with the ith element removed.

$$\begin{aligned} fold(X) &= X_i \circ fold(X_{\neg i}) && \text{By commutativity and associativity} \\ \frac{d\ fold(X)}{d\ X_i} &= \frac{d\ X_i \circ fold(X_{\neg i})}{d\ X_i} \\ &= D(X_i \circ fold(X_{\neg i}), X_i) && \text{By the precondition (eq. 1)} \\ &= D(fold(X), X_i) \end{aligned}$$

□

The most immediate counterexample to condition 1 is multiplication: $a \circ b = ab$. If $a = 0$, then $\frac{d\ ab}{d\ a} = b$ can not be expressed as a function of $ab = 0$ and $a = 0$. However, as we shall see, for many monoids of interest the condition does hold.

2.1 GeMMMapReduce

The benefits of not materializing partial values is most apparent in the case where the fold is applied to the results of matrix multiplications, one such instance is attention:

$$\begin{aligned} Q &: \mathbb{R}^{M \times F} \\ K &: \mathbb{R}^{N \times F} \\ V &: \mathbb{R}^{N \times D} \\ \text{attention}(Q, K, V) &= \text{softmax}(QK^\intercal)V \\ \text{attention}(Q, K, V)_i &= Y_{iv} \\ &\text{where} \\ Y_i &= \prod_j H_{ij} \\ H_{ij} &= \{z = Q_i K_j^\intercal,\ v = V_j\} \\ a \circ b &= \begin{Bmatrix} z = & \ln(e^{a_z} + e^{b_z}) \\ v = & a_v e^{a_z - z} + b_v e^{b_z - z} \end{Bmatrix} \end{aligned}$$

Here, we can think of H as an M by N matrix of T-values, with each such value consisting of an attention weight and value. Combining two such values results in a new weight, corresponding to a sum of their respective weights in log-space, and a new value corresponding to the weighted average of the two values. Since the result is a fold over the N-dimension, the full matrix does not have to be materialized. We can compute appropriate chunks of the matrix, fold the chunk, and add to a running total, only realizing the M normalizing factors, and the $M \times D$-sized weighted sum. Flash attention Dao et al. (2022) can be seen as an instance of this general approach.

We also have that condition 1 holds:

$$\begin{aligned} c &= a \circ b \\ \frac{d\, c}{d\, a} &= g \mapsto \begin{pmatrix} z: & (g_z + g_v^\intercal (a_v - c_v)) \exp(a_z - c_z) \\ v: & g_v \exp(a_z - c_z) \end{pmatrix} \\ &= D(c, a) \end{aligned}$$

Which due to theorem 1 implies that recomputation can be limited to the mapping from Q, K, V to H-values:

Listing 1: Attention example using elementwise MMMapReduce.

```
def forward(Q, K, V):
    """
    Q: M x F matrix
    K: N x F matrix
    V: N x D matrix
    returns:
    A: M-vector of T-values
    """
    A = full(id) # M-vector of identity T-values
    for (i, j) in (M x N):
        H = {z: Q[i] @ K[j], v: V[j]}
        A[i] = A[i] o H
    return A

def backward(A, Q, K, V, gA):
    """
    A: M-vector of T-values
    gA: output (A) gradient
    returns: gQ, gK, gV
    """
    gQ, gK, gV = [zeros_like(p) for p in [Q, K, V]]
    for (i, j) in (M x N):
        # local H and H-gradient.
        # H = recomputation of monoid elements.
        H = {z: Q[i] @ K[j], v: V[j]}
        gH = D(A[i], H)(gA[i])
        gQ[i] += gH.z * K[j]
        gK[j] += gH.z * Q[i]
        gV[j] += gH.v
    return gQ, gK, gV
```

In table 1 we give further examples of this approach applied to cross entropy against class indices, cross entropy between two distributions with logits given by matrix multiplication, and two-layer MLPs. For proofs of condition 1 for these, see Appendix A. The general pattern we observe is the following: We are given a tuple of input matrices (e.g. $\mathbf{X} = (Q, K, V)$), with corresponding shapes (e.g. $\mathbf{S} = (M \times F, N \times F, N \times D)$). The H-matrix (which we do not wish to materialize in full) has a shape derived from a subset of the input shapes $\mathbf{S}$ (e.g. $M \times N$). To compute the aggregate A we fold over some dimension(s) of the H-matrix, (e.g. along the N-dimension), which can be done iteratively,

only requiring us to realize arbitrarily small chunks (down to 1 element) of the H-matrix. The final result is then a function of the fold (e.g. $Y_i = A_{iv}$).

Input	Map	Reduce
$Q : \mathbb{R}^{M \times F}$ $K : \mathbb{R}^{N \times F}$ $V : \mathbb{R}^{N \times D}$ $\downarrow$ $Y : \mathbb{R}^{M \times D}$	$T : \{z : \mathbb{R}, v : \mathbb{R}^D\}$ $H : T^{M \times N}$ $H_{ij} = \begin{Bmatrix} z : & Q_i K_j^\intercal \\ v : & V_j \end{Bmatrix}$	$A : T^M$ $A_i = \prod_j H_{ij}$ $a \circ b = \begin{Bmatrix} z : & \ln(e^{a_z} + e^{b_z}) \\ v : & a_v e^{a_z - z} + b_v e^{b_z - z} \end{Bmatrix}$ $Y_i = A_{iv}$
Attention: $Y = \text{softmax}(QK^\intercal)V$		
$P : \mathbb{R}^{M \times D}$ $C : \mathbb{R}^{N \times D}$ $T : N^M$ $\downarrow$ $Y : \mathbb{R}^M$	$T : \{p : \mathbb{R}, n : \mathbb{R}\}$ $H : T^{M \times N}$ $H_{ij} = \begin{Bmatrix} p : & P_i C_j^\intercal \\ n : & T_i = j \ ? \ p \ : \ 0 \end{Bmatrix}$	$A : T^M$ $A_i = \prod_j H_{ij}$ $a \circ b = \begin{Bmatrix} p : & \ln(e^{a_p} + e^{b_p}) \\ n : & a_n + b_n \end{Bmatrix}$ $Y_i = A_{ip} - A_{in}$
Cross Entropy: $Y = \text{cross-entropy}(\text{logits} = PC^\intercal, \text{targets} = T)$		
$P^s : \mathbb{R}^{M \times D}$ $C^s : \mathbb{R}^{N \times D}$ $P^t : \mathbb{R}^{M \times E}$ $C^t : \mathbb{R}^{N \times E}$ $\downarrow$ $Y : \mathbb{R}^M$	$T : \{p : \mathbb{R}, q : \mathbb{R}, n : \mathbb{R}\}$ $H : T^{M \times N}$ $H_{ij} = \begin{Bmatrix} q : & P_i^s C_j^{s\intercal} \\ p : & P_i^t C_j^{t\intercal} \\ n : & q \end{Bmatrix}$	$A : T^M$ $A_i = \prod_j H_{ij}$ $a \circ b = \begin{Bmatrix} q : & \ln(e^{a_q} + e^{b_q}) \\ p : & \ln(e^{a_p} + e^{b_p}) \\ n : & a_n e^{a_p - p} + b_n e^{b_p - p} \end{Bmatrix}$ $Y_i = A_{iq} - A_{in}$
Cross Entropy: $Y = \text{cross-entropy}(\text{logits} = P^s C^{s\intercal}, \text{targets} = \text{softmax}(P^t C^{t\intercal}))$		
$X : \mathbb{R}^{B \times M}$ $P : \mathbb{R}^{M \times K}$ $Q : \mathbb{R}^{K \times N}$ $\downarrow$ $Y : \mathbb{R}^{B \times N}$	$T : \{v : \mathbb{R}^N\}$ $H : T^{B \times K}$ $H_{ij} = \{v : \sigma(X_i P_{:j}) Q_j\}$	$A : T^B$ $A_i = \prod_j X_{ij}$ $a \circ b = a_v + b_v$ $Y_i = A_{iv}$
MLP: $Y = \sigma(XP)Q$		

Table 1: Examples of Monoids T, map functions from inputs to intermediate H-values, reduction operations, and out projections that realizes different operations used in Artificial Neural Networks.

2.2 Batching and Compute efficiency

The pseudocode in listing 1 works elementwise, this is not really appropriate for efficient computation. A more GPU-friendly approach would be to compute the intermediate values and the gradient over appropriately sized slices of the H-matrix, while simultaneously folding it along the fold dimension, as is done in listing 2, where we show a general proof of concept implementation of the GeMMMapReduce operation, given user-provided functions `proj_fold`, `proj_fold_bwd`, `binary_reduce`, and helper functions `chunker`, `init`.

Listing 2: Generalized, Batched, GeMMMapReduce example, implemented as torch function.

```
class GeMMMapReduce(torch.autograd.Function):
    @staticmethod
    def forward(*X):
        A = init(X)
        for aslice, xslice in chunker(X):
            a = aslice(A)
            x = xslice(X)
            local_a = proj_fold(x)
            new_a = binary_reduce(a, local_a)
            for view, new_val in zip(a, new_a):
                view.copy_(new_val)
        return A

    @staticmethod
    def setup_context(ctx, inputs, outputs):
        ctx.num_inputs = len(inputs)
        ctx.save_for_backward(*inputs, *outputs)

    @staticmethod
    @once_differentiable
    def backward(ctx, *gA):
        X = ctx.saved_tensors[:ctx.num_inputs]
        A = ctx.saved_tensors[ctx.num_inputs:]
        gX = [p.new_zeros(p.shape) for p in X]
        for aslice, xslice in chunker(X):
            # Extract chunks
            x = xslice(X)
            a = aslice(A)
            ga = aslice(gA)
            # recompute and calculate local gradients.
            gx = proj_fold_bwd(x, a, ga)
            # Add local gradients to global.
            for g, d in zip(xslice(gX), gx):
                g.add_(d)
        return tuple(gX)
```

3 Discussion

In general, the benefit of activation recomputation is to reduce memory at the cost of the extra compute needed for recomputation. This obviously holds for

GeMMMapReduce as well. If the amount of memory saved is inconsequential, or is dwarfed by the cost of recomputation, there is little benefit of the approach.

The hope is that with a general and *simple* framework that enables memory and cache-efficient layers, we might be able to iterate and innovate faster by making memory efficient layers by design.

A Gradients for CrossEntropy and MLP

For the examples in table 1 we also have that condition 1 holds, and by extension that theorem 1 holds:

Monoid

$$\begin{aligned} T &: \{p : \mathbb{R}, n : \mathbb{R}\} \\ a \circ b &= \begin{Bmatrix} p : & \ln(e^{a_p} + e^{b_p}) \\ n : & a_n + b_n \end{Bmatrix} \\ &= c \end{aligned}$$

Gradient

$$\frac{d\ c}{d\ a} = g \mapsto \begin{Bmatrix} p : & g_p e^{a_p - c_p} \\ n : & g_n \end{Bmatrix}$$

CrossEntropy

$$\begin{aligned} T &: \{q : \mathbb{R}, p : \mathbb{R}, n : \mathbb{R}\} \\ a \circ b &= \begin{Bmatrix} q : & \ln(e^{a_q} + e^{b_q}) \\ p : & \ln(e^{a_p} + e^{b_p}) \\ n : & a_n e^{a_p - p} + b_n e^{b_p - p} \end{Bmatrix} \\ &= c \end{aligned}$$

$$\frac{d\ c}{d\ a} = g \mapsto \begin{Bmatrix} q : & g_q e^{a_q - c_q} \\ p : & (g_p + g_n(a_n - c_n))e^{a_p - c_p} \\ n : & g_n e^{a_p - c_p} \end{Bmatrix}$$

CrossEntropy between two distributions

$$\begin{aligned} T &: \{v : \mathbb{R}^D\} \\ a \circ b &= \{v : a_v + b_v\} \\ &= c \end{aligned}$$

$$\frac{d\ c}{d\ a} = g \mapsto g$$

MLP

References

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Re, C. (2022). Flashattention: Fast and memory-efficient exact attention with IO-awareness. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.