Memory and IO in deep learning

Lecture 12 — CS 577 Deep Learning

Instructor: Yutong Wang

Computer Science
Illinois Institute of Technology

November 6, 2024

Lec 13 is a seminar

Memory and IO in deep learning

Lecture 12 — CS 577 Deep Learning

So far...

- Backprop
- 2D Convol
  lower to Linear Alg
- Parallelize seq model
  RNN → Transformer

Instructor: Yutong Wang

Computer Science
Illinois Institute of Technology

November 6, 2024

No discussion of memory so far

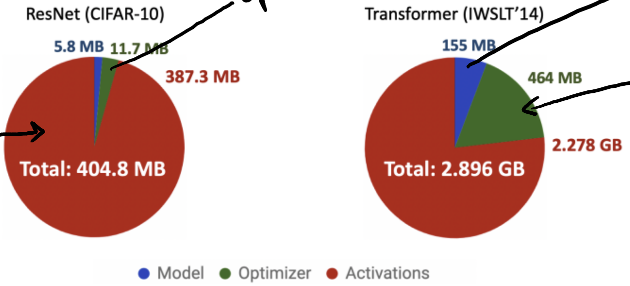
The memory wall

Memory usage breakdown

optimizer has "states" - stateful

- Model
- Training produces

"Activation"
Intermediate



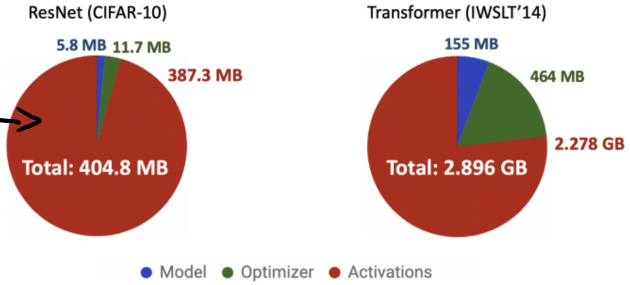
Weights

Optimizer
- SGD no overhead
- Momentum keep track of history

From Sohoni et al. 2019 "Low-memory neural network training: A technical report"

The memory wall

Biggest piece of pie

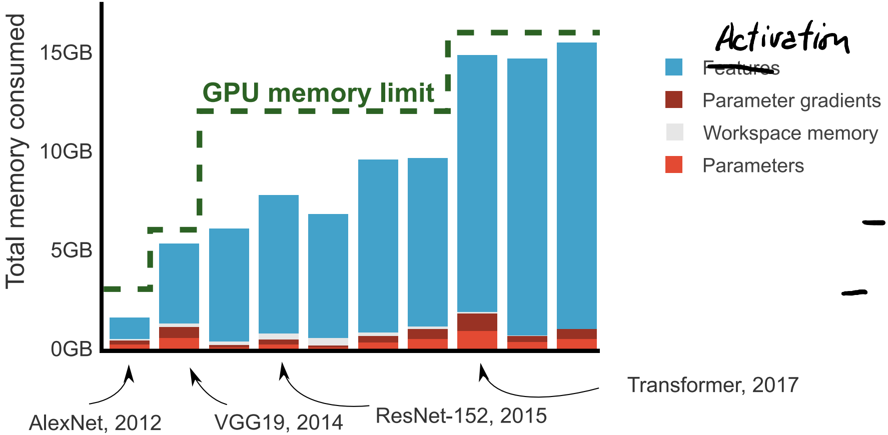


From Sohoni et al. 2019 “Low-memory neural network training: A technical report”

The memory wall

DL practitioners fill up the memory capacity

- Big batch
- Big model



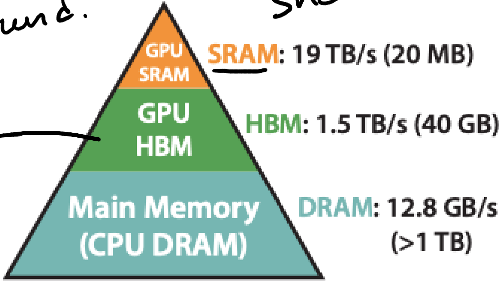
From Jain et al. 2020 “Checkmate: Breaking the memory wall with optimal tensor rematerialization”

Flash attention

IO input-output
moving data around.

shared memory on chip
physical close to compute

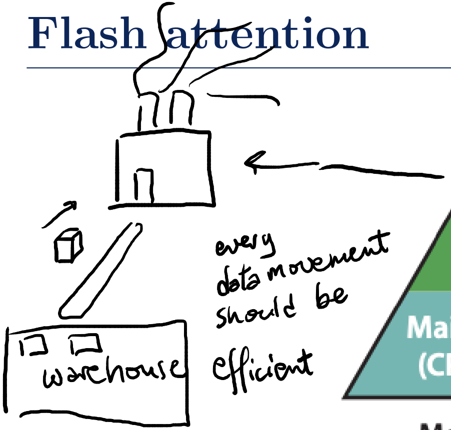
"off-chip"
more space but slow



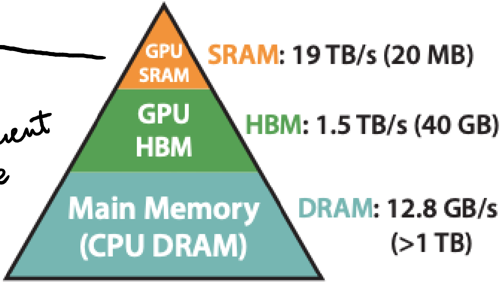
Memory Hierarchy with Bandwidth & Memory Size

From Dao et al. 2022 "FlashAttention: Fast and memory-efficient exact attention with io-awareness"

Flash attention



every data movement should be efficient



Memory Hierarchy with Bandwidth & Memory Size

From Dao et al. 2022 “FlashAttention: Fast and memory-efficient exact attention with io-awareness”

Memory and IO

- DL workloads often memory limited
- Moving data (IO) is costly

Roadmap for today

pytorch

tensor(___ , requires_grad)

Memory:

- Warm-up: `requires_grad` ← discard grad? tool for improving memory
- Gradient checkpointing

rematerialization ← discard activ?

IO:

- FlashAttention: how to compute the attention module without moving around so much data

computing standard attention

Next: `requires_grad`

Size of NumPy array

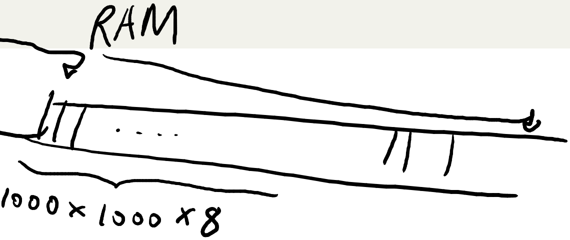
```python
import numpy as np
import sys

n = 1000
X_np = np.random.randn(n,n)
sys.getsizeof(X_np)

# >>> 8000128      # <- unit in bytes
```

Array in C

```c
double *X;

X = (double *)malloc(n * n * sizeof(double));
```

```
n*n*8

# >>> 8000000
```

≈ 8 Mb



8 bytes
in 1 float64
double

Size of NumPy array

```python
import numpy as np
import sys

n = 1000
X_np = np.random.randn(n,n)
sys.getsizeof(X_np)

# >>> 8000128    # <--- extra bytes for metadata
```

```python
n*n*8      # <--- size_of_float64  times  num_of_elements_in_X_np

# >>> 8000000
```

Aside: Size of PyTorch array

```python
import torch

n = 1000
X_torch = torch.randn(n,n)
sys.getsizeof(X_torch.untyped_storage())

# >>> 4000072
```

reference

```python
n*n*4      # <--- size_of_float32  times  num_of_elements_in_X_torch

# >>> 4000000
```

bfloat16 – google brain

– Quantization: save memory by making things less precise

tracemalloc

Allows tracking only memory allocated by numpy

```
import tracemalloc

# [...]
print("Simple example")

start_trace()

a = 1.0*np.zeros((n,n))      # n = 1000

print_trace_stats(snapshot_trace())
end_trace()
# >>> Simple example
# >>> memory allocated: 8 MB
```

track memory allocated by a module numpy

$$\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}$$

$1000 \times 1000 \times 8$

tracemalloc

Allows tracking only memory allocated by numpy

```
import tracemalloc

# [...]
print("No discarding")
start_trace()

a = 1.0*np.zeros((n,n))      # n = 1000
b = a*a
c = np.sum(b,axis=1)

print_trace_stats(snapshot_trace())
end_trace()
# >>> No discarding
# >>> memory allocated: 16 MB
```

8MB

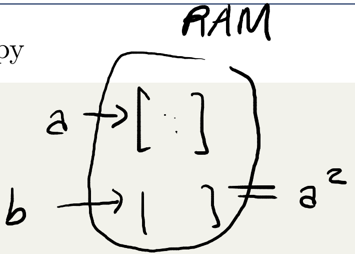
8MB

1000×8 8 kb

16.008

tracemalloc

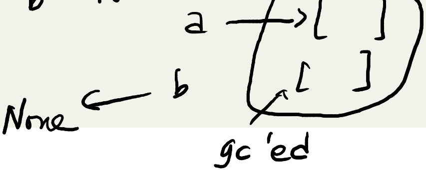
Allows tracking only memory allocated by numpy



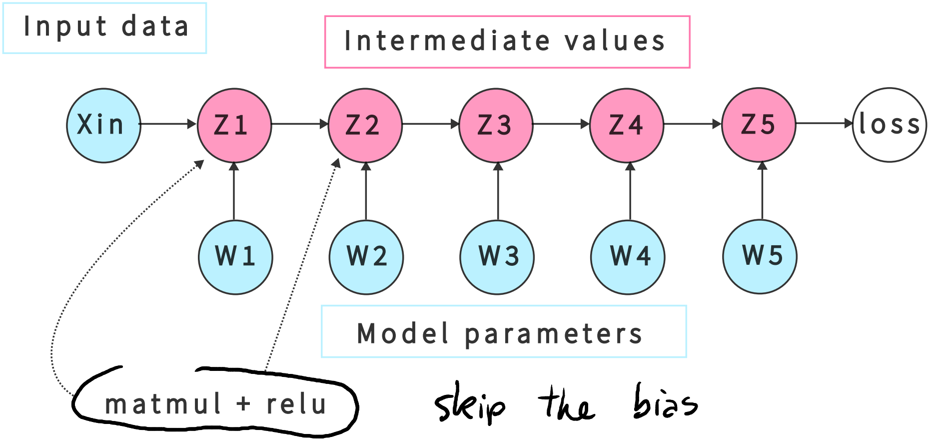

```
import tracemalloc

# [...]
print("Discarding")
start_trace()

a = 1.0*np.zeros((n,n))     # n = 1000
b = a*a
c = np.sum(b,axis=1)
b = None

print_trace_stats(snapshot_trace())
end_trace()
# >>> With discarding
# >>> memory allocated: 8 MB
```

← want to free up the space.

b = None



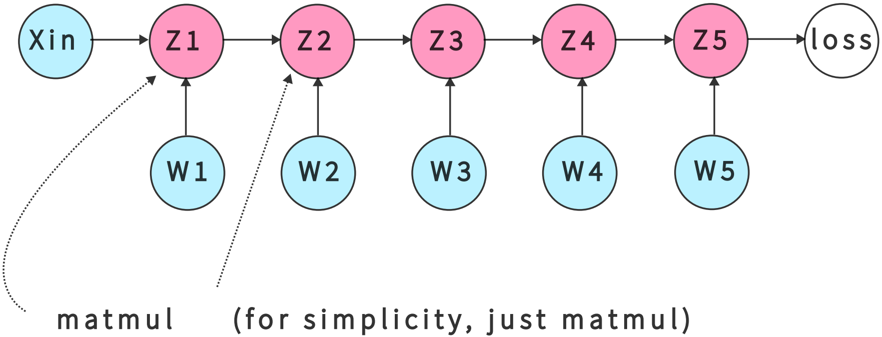
What is needed to calculate the backward?



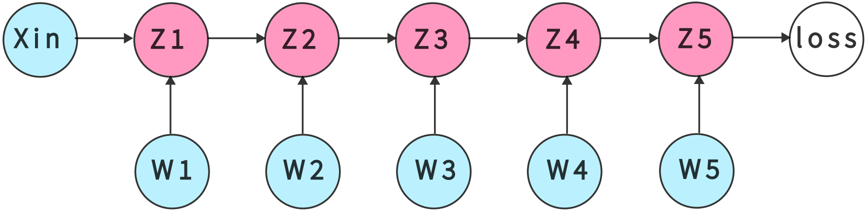
skip the bias

What is needed to calculate the backward?

Not a good model, but simple enough for memory analysis



What is needed to calculate the backward?

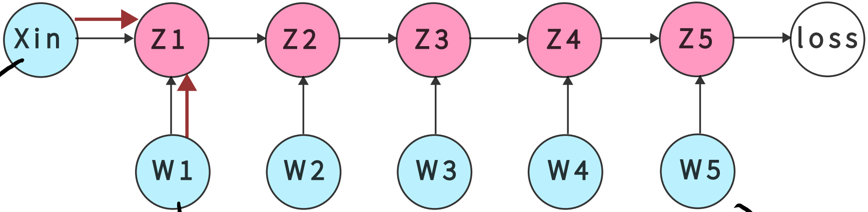


What is needed to calculate the backward?

$Z1.shape = [\underbrace{4000}_{\text{batch\_size}}, 1000]$

32 MB

```
Z1 = ag.matmul(Xin, W1)
```



shape
(N, d)
[4000, 1000]

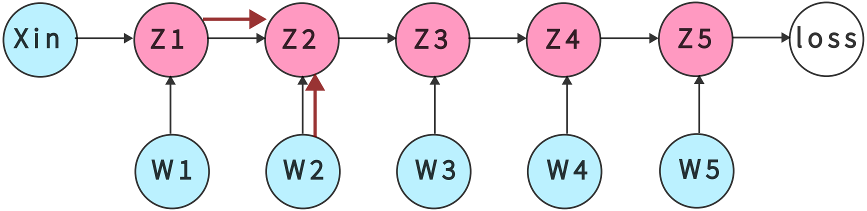
[1000, 1000]

W5.value has 8 MB

What is needed to calculate the backward?

Z2.value — 32 MB
.grad — 32 MB
total 64 MB
per
layer

```
Z2 = ag.matmul(Z1, W2)
```

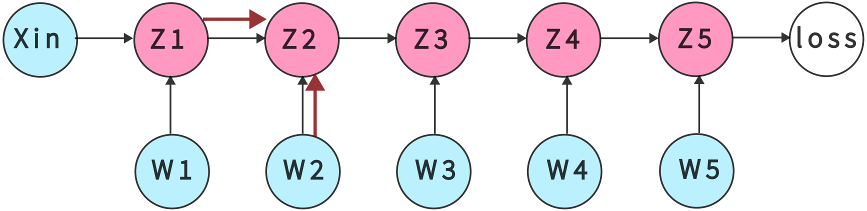


What is needed to calculate the backward?

(Inside constructor for Z2)...

```
self.value = 1.0*value
self.grad = np.zeros_like(self.value)
```

wasted

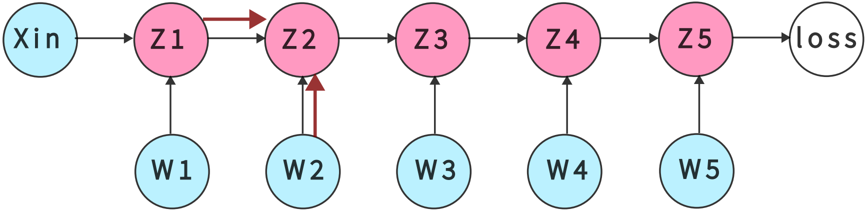


Array sizes!

(Inside constructor for Z2)...

```
self.value = 1.0*value
self.grad = np.zeros_like(self.value)
```

64 MB



What is needed to calculate the backward?

```
def forward(x, weights):
    for w in weights:
        x = ag.matmul(x, w)
    return ag.sum(x)
```

What is needed to calculate the backward?

```
def forward_traced(x, weights):
    start_trace()                    # tracing
    mem_usage = []                   # tracing
    for w in weights:
        x = ag.matmul(x, w)
        mem_usage.append(snapshot_trace()) # tracing
    l = ag.sum(x)
    mem_usage.append(snapshot_trace()) # tracing
    end_trace() # tracing
    return l, mem_usage
```

64 MB

128 MB

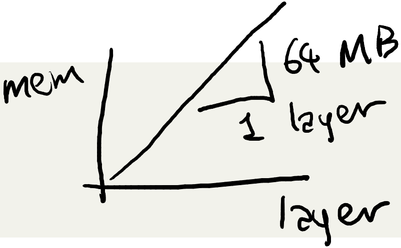
⋮

What is needed to calculate the backward?

```python
for i, trace_stats in enumerate(mem_usage_forward):
    print(f"layer {i}")
    print_trace_stats(trace_stats)
```

Output:

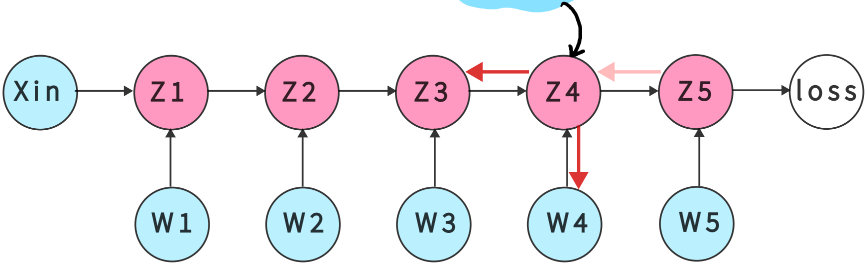
```
layer 0
memory allocated: 64 MB
layer 1
memory allocated: 128 MB
layer 2
memory allocated: 192 MB
```



Observation: double the amount of memory actually needed!

What is needed to calculate the backward?

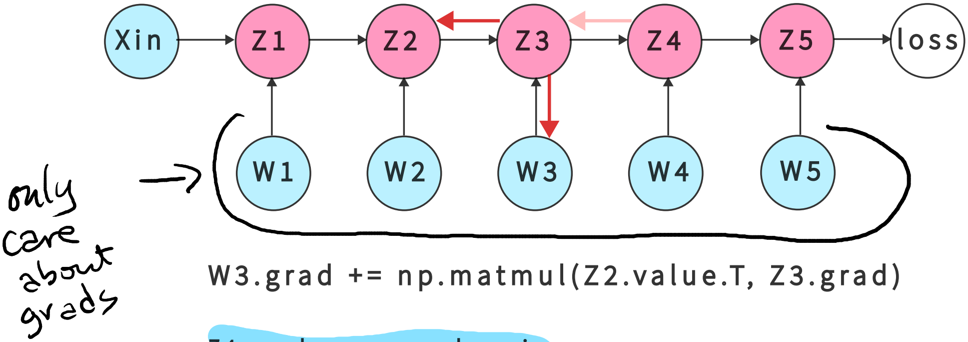
```
Z3.grad += np.matmul(Z4.grad, W4.value.T)
```




```
W4.grad += np.matmul(Z3.value.T, Z4.grad)
```

What is needed to calculate the backward?

```
Z2.grad += np.matmul(Z3.grad, W3.value.T)
```




```
W3.grad += np.matmul(Z2.value.T, Z3.grad)
```

Z4.grad never used again

Backward

```python
    def backward(self):
        self.grad = np.array(1.0)

        topo_order = self.topological_sort()

        for node in reversed(topo_order):
            node._backward()
        return None
```

Backward with tracing

```
def backward(self):
    self.grad = np.array(1.0)

    topo_order = self.topological_sort()

    start_trace()  # tracing
    mem_usage = [] # tracing

    for node in reversed(topo_order):
        node._backward()
        mem_usage.append(snapshot_trace()) # tracing
    end_trace() # tracing
    return mem_usage
```

Backward with tracing

```python
for i, trace_stats in enumerate(mem_usage_backward):
    print(f"backward step {i}")
    print_trace_stats(trace_stats)
```

Output:

```
backward step 0
memory allocated: 0 MB
backward step 1
memory allocated: 0 MB
backward step 2
memory allocated: 0 MB
```

allocated everything already

Observation: After calling backward on a tensor, its grad is no longer needed

Backward with tracing

```python
for i, trace_stats in enumerate(mem_usage_backward):
    print(f"backward step {i}")
    print_trace_stats(trace_stats)
```

Output:

```
backward step 0
memory allocated: 0 MB
backward step 1
memory allocated: 0 MB
backward step 2
memory allocated: 0 MB
```

Observation: After calling backward on a tensor, its grad is no longer needed

Next: Shift the task of initialize the grad to the backward pass

Keeping track of grad — is it needed?

Constructor

```
class Tensor: # Tensor with grads
    def __init__(self,
                 value,
        # [...]

        self.value = 1.0*value
        self.grad = np.zeros_like(self.value) # <---- WASTEFUL!
```

Keeping track of grad — is it needed?

```
class Tensor: # Tensor with grads
    def __init__(self,
                 value,
                 requires_grad=False, # <-- Flag for keeping grad
        # [...]

        self.value = 1.0*value

        self.grad = None
        if self.requires_grad:
            self.grad = np.zeros_like(self.value)
```

Keeping track of grad — is it needed?

```
    class Tensor: # Tensor with grads
        def __init__(self,
                     value,
                     requires_grad=False, # <-- Flag for keeping grad
            # [...]

            self.value = 1.0*value

            self.grad = None
            if self.requires_grad:
                self.grad = np.zeros_like(self.value)
# [...]

weights = [ag.Tensor(0.02*np.random.randn(dim_hidden, dim_hidden),
                     requires_grad = True) for _ in range(num_layers)]
#                    ^^^^^^^^^^^^^^^^^^^^
```

Keeping track of grad — is it needed?

```
    class Tensor: # Tensor with grads
        def __init__(self,
                     value,
                     requires_grad=False, # <-- Flag for keeping grad
            # [...]

            self.value = 1.0*value

            self.grad = None
            if self.requires_grad:
                self.grad = np.zeros_like(self.value)
# [...]

    for w in weights:
        x = ag.matmul(x, w)
        # x.requires_grad == False by default
```

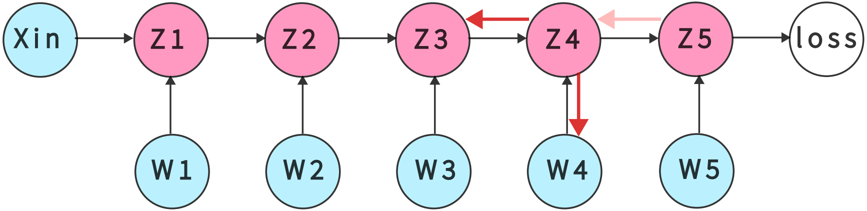
Problem: during backward, we might encounter `None`'s when we should expect `grad`.

What is needed to calculate the backward?

is None

assume somehow we initialized this

```
Z3.grad += np.matmul(Z4.grad, W4.value.T)
```




```
W4.grad += np.matmul(Z3.value.T, Z4.grad)
```

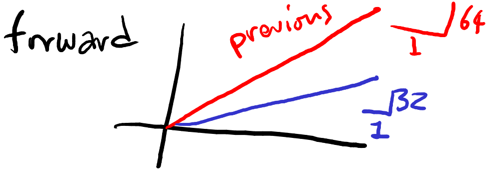
this is fine

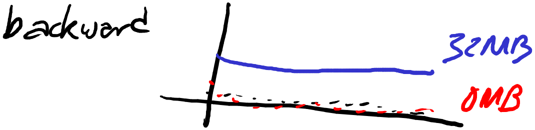
Exercise 1: discarding the grad

and def sum(·); _backward

```python
# inside ag.__matmul__
            def _backward():
                # YOUR CODE HERE FOR initializing the grad

                self.grad += np.matmul(output.grad, other.value.T)
                other.grad += np.matmul(self.value.T, output.grad)

                # YOUR CODE HERE FOR discarding the grad
                return None
```

If your answer is correct, then the "sanity checks" should have the expected output.



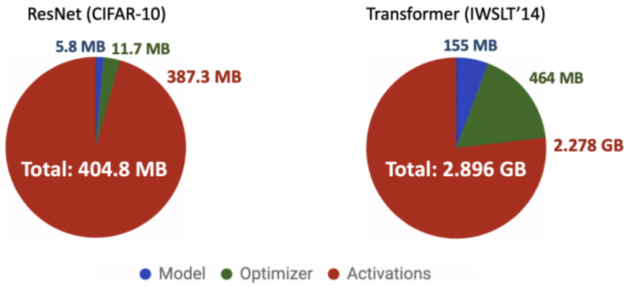


The memory wall

- We've reduced the memory consumption of forward by a factor 2 without incurring much additional backward memory burden.

The memory wall

- We've reduced the memory consumption of forward by a factor 2 without incurring much additional backward memory burden.
- But recall that...



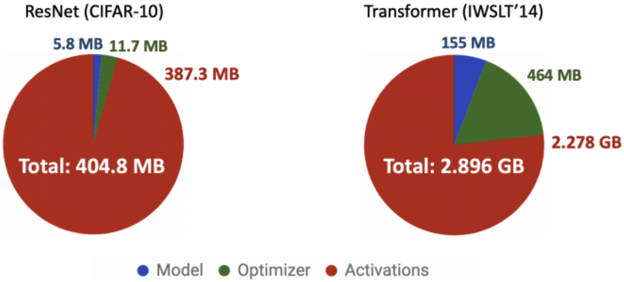
From Sohoni et al. 2019 "Low-memory neural network training: A technical report"

The memory wall

- We've reduced the memory consumption of forward by a factor 2 without incurring much additional backward memory burden.

The memory wall

- We've reduced the memory consumption of forward by a factor 2 without incurring much additional backward memory burden.
- But recall that...



From Sohoni et al. 2019 "Low-memory neural network training: A technical report"

The memory wall

Can we do something about this?

```
layer 0
memory allocated: 32 MB
layer 1
memory allocated: 65 MB
layer 2
memory allocated: 98 MB
layer 3
memory allocated: 131 MB
layer 4
memory allocated: 163 MB
...
```

Tensor Rematerialization

- Discard the activations as soon as you can

if all outgoing edge have been compute can delete (for forward pass)

Tensor Rematerialization

- Discard the activations as soon as you can
- "Beam" them back/recompute them only when you need them

From Berman et al. 1995 "Star Trek: Voyager"

Tensor Rematerialization

input

recompute

!

back

loss

- Discard the activations as soon as you can
- "Beam" them back/recompute them only when you need them

From Berman et al. 1995 "Star Trek: Voyager"

50th layer

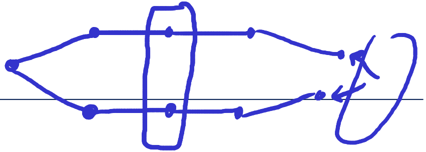
don't discard

- Save *some* activations as "checkpoints" to speed things up

only recompute from checkpoint

100 depth

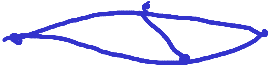
Tensor Rematerialization



- Discard the activations as soon as you can
- "Beam" them back/recompute them only when you need them

Mix of Experts

From Berman et al. 1995 "Star Trek: Voyager"

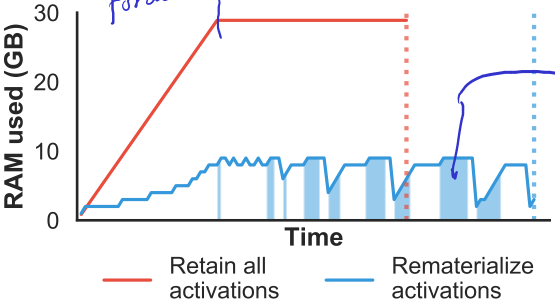


- Save *some* activations as "checkpoints" to speed things up

$h^{(1)} \to h^{(2)} \to h^{(3)}$

Tensor Rematerialization

forward backward



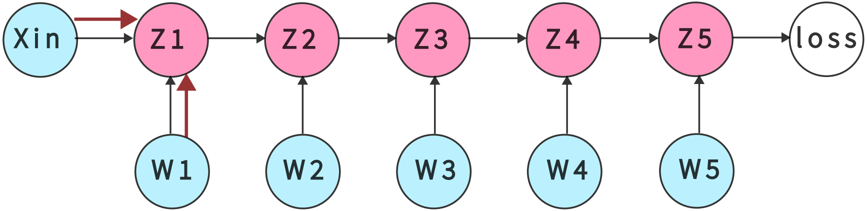
rematerialization

From Jain et al. 2020 “Checkmate: Breaking the memory wall with optimal tensor rematerialization”

- Also known as **gradient checkpoint**, activation checkpointing, recomputation...

Rematerialization: forward phase

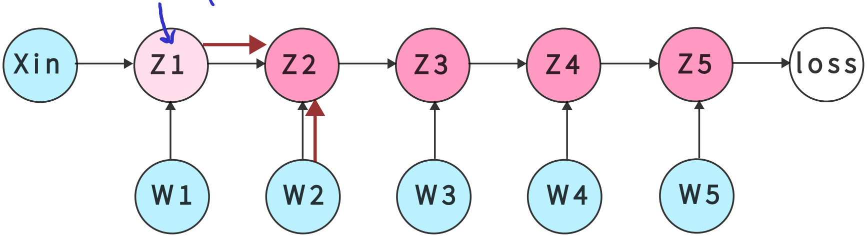
```
Z1 = ag.matmul(Xin, W1)
```



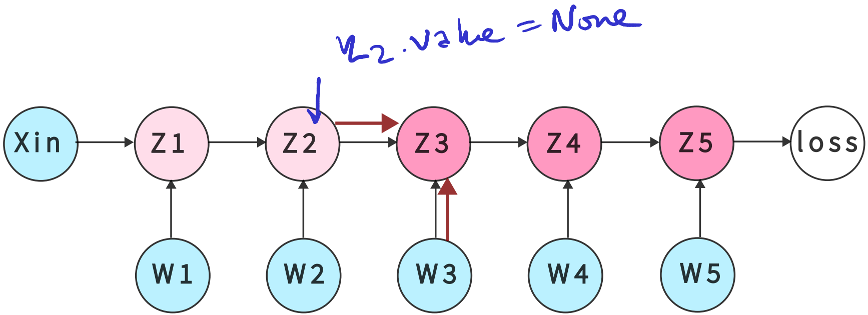
Rematerialization: forward phase

```
Z2 = ag.matmul(Z1, W2)
```

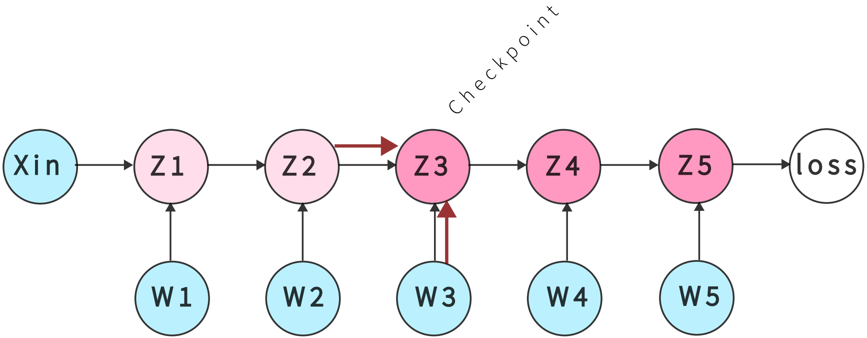
Z_1.value = None



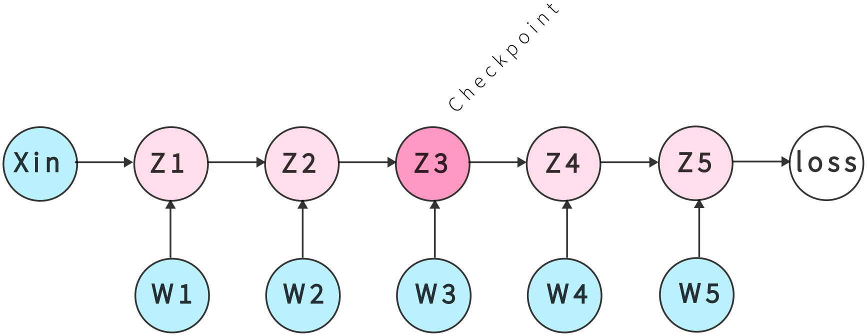
Rematerialization: forward phase



Rematerialization: forward phase

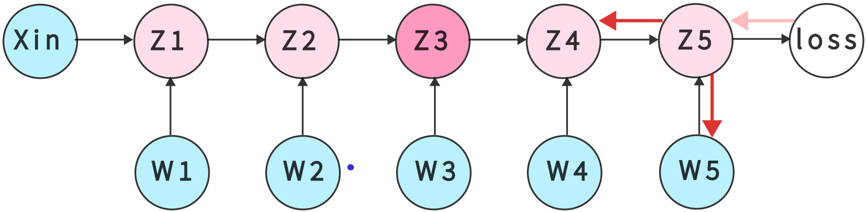


Rematerialization: forward phase



Rematerialization: backward phase

```
Z4.grad += np.matmul(Z5.grad, W5.value.T)
```

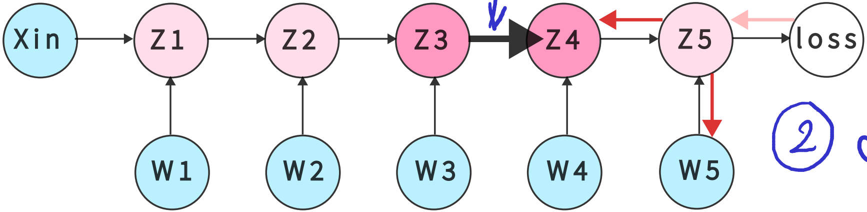



```
W5.grad += np.matmul(Z4.value.T, Z5.grad)
```

Rematerialization: backward phase

1 have the ability to compute partial forward

```
Z4.grad += np.matmul(Z5.grad, W5.value.T)
```

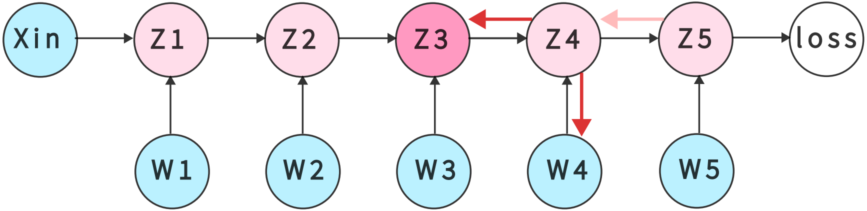


2 delete the intermediate values.

```
W5.grad += np.matmul(Z4.value.T, Z5.grad)
```

Rematerialization: backward phase

```
Z3.grad += np.matmul(Z4.grad, W4.value.T)
```




```
W4.grad += np.matmul(Z3.value.T, Z4.grad)
```

Exercise 2: rematerialization

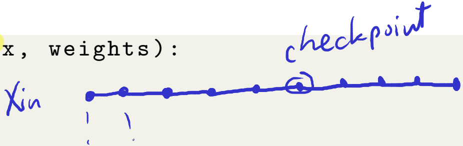
```
class Tensor: # Tensor with grads
    def __init__(self,
                 value,
                 requires_grad=False,
                 rematerializer = None, # None means don't rematerialize, KEEP
# [...]
```

partial forward

pass information to each node a function to recompute itself.

None means checkpoint this tensor no discard.

Exercise 2: rematerialization




```
def forward_traced_with_rematerializer(x, weights):
    start_trace()
    mem_usage = [] # tracing

    checkpoints = [5] # these layers are checkpoints

    farthest_checkpoint = 0 # this is the input data
    x_at_farthest_checkpoint = x

    for i, w in enumerate(weights): # main training loop
        if i in checkpoints:
            x = ag.matmul(x, w)
            farthest_checkpoint = i
            x_at_farthest_checkpoint = x
# [...]
```

rematerializer = None

update check

keep pointer alive

Exercise 2: rematerialization

ag ⟷ pytorch

numpy ⟷ ATEN

```
for i, w in enumerate(weights): # main training loop
    if i in checkpoints:
        x = ag.matmul(x, w)
        farthest_checkpoint = i
        x_at_farthest_checkpoint = x
    else:
        def _rematerializer():
            xval = x_at_farthest_checkpoint.value
            for w in weights[farthest_checkpoint:(i+1)]:
                xval = np.matmul(xval,w.value)
            return xval
        x = ag.matmul(x, w)
        x.rematerializer = _rematerializer
```

not checkpoint

do with plain numpy to avoid Autograd

recipe for reconstructing x

Exercise 2: task 1: rematerialize!

```python
    def backward(self):
        # [...]
        for node in reversed(topo_order):
            # YOUR CODE HERE FOR rematerializing "input.value", if it is none

            node._backward()
```

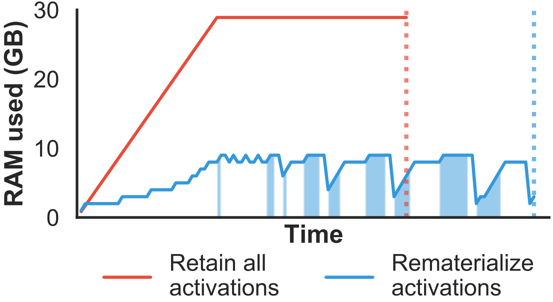
should see inputs whose values are rematerialization

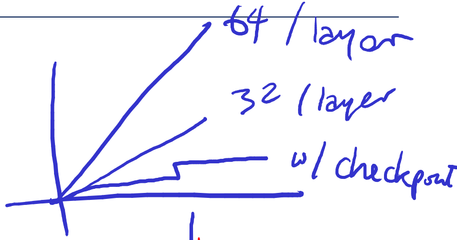
Exercise 2: task 2: discard!

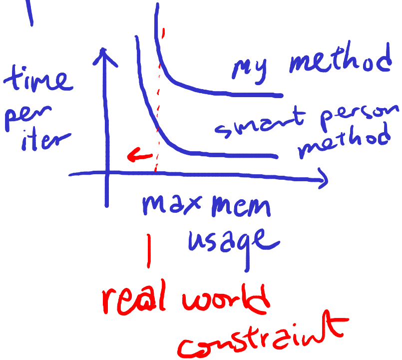
```python
        def __matmul__(self,other):
            # [...]
            def _backward():
                # [...]
                self.grad += np.matmul(output.grad, other.value.T)
                other.grad += np.matmul(self.value.T, output.grad)

                # YOUR CODE HERE FOR discarding activations for "self" and "other"
                # hint: add a helper function to make it neater
                # hint: see "discard_value_if_has_rematerializer" below
                # [lines skipped ...]
                return None

            output._backward = _backward
            # YOUR CODE HERE FOR discarding activations for "self" and "other"
```

Summary

- Less memory used in the forward

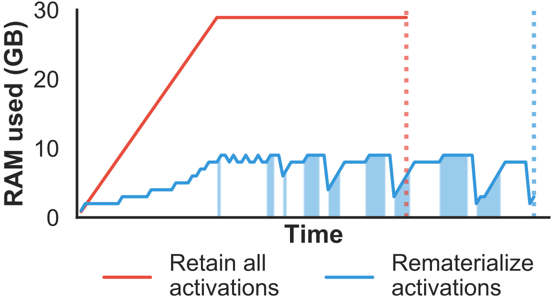






Summary

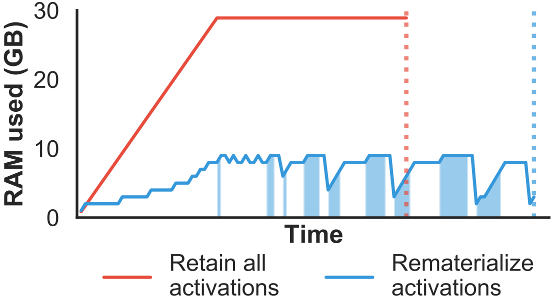
- Less memory used in the forward



- More computation (for rematerialization).
  (75 matmuls w/ rematerialization vs. 30 matmuls w/o rematerialization).

Summary

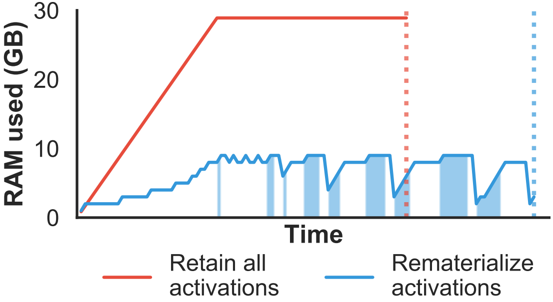
- Less memory used in the forward



- More computation (for rematerialization).
  (75 matmuls w/ rematerialization vs. 30 matmuls w/o rematerialization).
- **Question**: what about more complicated computational graph?

Summary

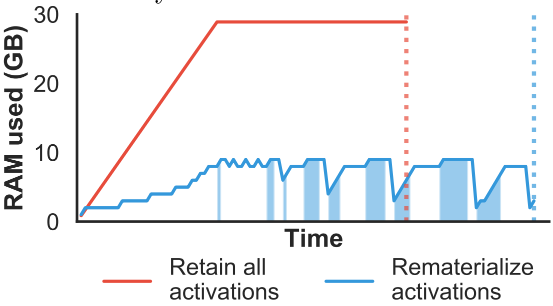
- Less memory used in the forward



- More computation (for rematerialization).
  (75 matmuls w/ rematerialization vs. 30 matmuls w/o rematerialization).
- **Question**: what about more complicated computational graph?
- **Question**: optimal way of selecting the checkpoints?

Summary

- Less memory used in the forward



- More computation (for rematerialization).
  (75 matmuls w/ rematerialization vs. 30 matmuls w/o rematerialization).
- **Question**: what about more complicated computational graph?
- **Question**: optimal way of selecting the checkpoints?
- **Next**: IO and FlashAttention

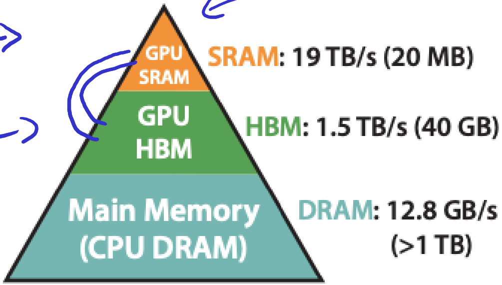
Flash attention

fast close to the chip

20 MB

40GB

High bandwidth memory



CPU RAM

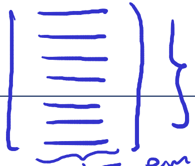
Hard drive

to SRAM is expensive

Memory Hierarchy with Bandwidth & Memory Size

From Dao et al. 2022 "FlashAttention: Fast and memory-efficient exact attention with io-awareness"

Self-attention

$X =$ 

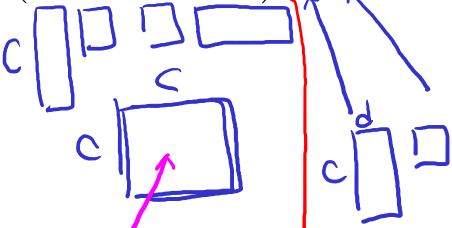
C context   Seq_len

left most dimension

d   embedding

- Let $\mathbf{X} \in \mathbb{R}^{C \times d}$ (note everything is transposed compared to last time)

$$\mathtt{attention}(\mathbf{X};\theta) := \mathrm{softmax}\left(\mathbf{X}\mathbf{W}^{(Q)}\mathbf{W}^{(K)\top}\mathbf{X}^{\top}\right)\mathbf{X}\mathbf{W}^{(V)} \in \mathbb{R}^{d \times C}.$$



C = want this to be very large (more tokens)

d = (moderately large) ~ 700

Huge 1,000,000 = C >> d = 700

Self-attention

- Let $\mathbf{X} \in \mathbb{R}^{C\times d}$ (note everything is transposed compared to last time)

$$\texttt{attention}(\mathbf{X};\theta) := \mathrm{softmax}\left(\mathbf{X}\mathbf{W}^{(Q)}\mathbf{W}^{(K)\top}\mathbf{X}^\top\right)\mathbf{X}\mathbf{W}^{(V)} \in \mathbb{R}^{d\times C}.$$

- parameters

$$\theta^{(\texttt{att})} = [\mathbf{W}^{(Q)}, \mathbf{W}^{(K)}, \mathbf{W}^{(V)}]$$

Self-attention

- Let $\mathbf{X} \in \mathbb{R}^{C \times d}$ (note everything is transposed compared to last time)

$$\mathtt{attention}(\mathbf{X}; \theta) := \operatorname{softmax}\left(\mathbf{X}\mathbf{W}^{(Q)}\mathbf{W}^{(K)\top}\mathbf{X}^{\top}\right)\mathbf{X}\mathbf{W}^{(V)} \in \mathbb{R}^{d \times C}.$$

- parameters

$$\theta^{(\mathtt{att})} = [\mathbf{W}^{(Q)}, \mathbf{W}^{(K)}, \mathbf{W}^{(V)}]$$

- ("Q, K, V" stands for "query", "key", "value", respectively) where

$$\mathbf{W}^{(Q)} \text{ and } \mathbf{W}^{(K)} \text{ and } \mathbf{W}^{(V)} \in \mathbb{R}^{d \times d}.$$

Self-attention

- Let $\mathbf{X} \in \mathbb{R}^{C\times d}$ (note everything is transposed compared to last time)

$$\texttt{attention}(\mathbf{X};\theta) := \operatorname{softmax}\left(\mathbf{X}\mathbf{W}^{(Q)}\mathbf{W}^{(K)\top}\mathbf{X}^\top\right)\mathbf{X}\mathbf{W}^{(V)} \in \mathbb{R}^{\cancel{d\times C}}.$$

$\mathbb{R}^{C\times d}$

- parameters

$$\theta^{(\texttt{att})} = [\mathbf{W}^{(Q)}, \mathbf{W}^{(K)}, \mathbf{W}^{(V)}]$$

- ("Q, K, V" stands for "query", "key", "value", respectively) where

$$\mathbf{W}^{(Q)} \text{ and } \mathbf{W}^{(K)} \text{ and } \mathbf{W}^{(V)} \in \mathbb{R}^{d\times d}.$$

$\mathbb{R}^{C\times d}$

- "**seq-2-seq**": maps the sequence $\mathbf{X}$ to another sequence $\texttt{attention}(\mathbf{X};\theta)$

Self-attention

- Let $\mathbf{X} \in \mathbb{R}^{C \times d}$ (note everything is transposed compared to last time)

$$\texttt{attention}(\mathbf{X}; \theta) := \mathrm{softmax}\Big(\underbrace{\mathbf{X}\mathbf{W}^{(Q)}}_{\mathbf{Q}}\underbrace{\mathbf{W}^{(K)\top}\mathbf{X}^\top}_{\mathbf{K}^\top}\Big)\underbrace{\mathbf{X}\mathbf{W}^{(V)}}_{\mathbf{V}} \in \mathbb{R}^{d \times C}.$$

- parameters

$$\theta^{(\texttt{att})} = [\mathbf{W}^{(Q)}, \mathbf{W}^{(K)}, \mathbf{W}^{(V)}]$$

- ("Q, K, V" stands for "query", "key", "value", respectively) where

$$\mathbf{W}^{(Q)} \text{ and } \mathbf{W}^{(K)} \text{ and } \mathbf{W}^{(V)} \in \mathbb{R}^{d \times d}.$$

- "**seq-2-seq**": maps the sequence $\mathbf{X}$ to another sequence $\texttt{attention}(\mathbf{X}; \theta)$

Self-attention

- Let $\mathbf{X} \in \mathbb{R}^{C \times d}$ (note everything is transposed compared to last time)

$$\texttt{attention}(\mathbf{X};\theta) := \text{softmax}\Big(\underbrace{\mathbf{Q}\mathbf{K}^\top}_{\mathbf{S}}\Big)\mathbf{V} \in \mathbb{R}^{d \times C}.$$

- parameters

$$\theta^{(\texttt{att})} = [\mathbf{W}^{(Q)}, \mathbf{W}^{(K)}, \mathbf{W}^{(V)}]$$

- ("Q, K, V" stands for "query", "key", "value", respectively)
  where

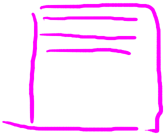
$$\mathbf{W}^{(Q)} \text{ and } \mathbf{W}^{(K)} \text{ and } \mathbf{W}^{(V)} \in \mathbb{R}^{d \times d}.$$

- "**seq-2-seq**": maps the sequence $\mathbf{X}$ to another sequence $\texttt{attention}(\mathbf{X};\theta)$

Self-attention

- Let $\mathbf{X} \in \mathbb{R}^{C \times d}$ (note everything is transposed compared to last time)

$$\texttt{attention}(\mathbf{X}; \theta) := \underbrace{\mathrm{softmax}\Big(\underbrace{\mathbf{Q}\mathbf{K}^\top}_{\mathbf{S}}\Big)}_{\mathbf{P}=\mathrm{softmax}(\mathbf{S})} \mathbf{V} \in \mathbb{R}^{d \times C}.$$



row stochastic

- parameters

$$\theta^{(\texttt{att})} = [\mathbf{W}^{(Q)}, \mathbf{W}^{(K)}, \mathbf{W}^{(V)}]$$

- ("Q, K, V" stands for "query", "key", "value", respectively) where

$$\mathbf{W}^{(Q)} \text{ and } \mathbf{W}^{(K)} \text{ and } \mathbf{W}^{(V)} \in \mathbb{R}^{d \times d}.$$

- "**seq-2-seq**": maps the sequence $\mathbf{X}$ to another sequence $\texttt{attention}(\mathbf{X}; \theta)$

Self-attention

- Let $\mathbf{X} \in \mathbb{R}^{C\times d}$ (note everything is transposed compared to last time)

$$\texttt{attention}(\mathbf{X};\theta) := \text{softmax}\left(\mathbf{Q}\mathbf{K}^\top\right)\mathbf{V} \in \mathbb{R}^{d\times C}.$$

- parameters

$$\theta^{(\texttt{att})} = [\mathbf{W}^{(Q)}, \mathbf{W}^{(K)}, \mathbf{W}^{(V)}]$$

- ("Q, K, V" stands for "query", "key", "value", respectively) where

$$\mathbf{W}^{(Q)} \text{ and } \mathbf{W}^{(K)} \text{ and } \mathbf{W}^{(V)} \in \mathbb{R}^{d\times d}.$$

- "**seq-2-seq**": maps the sequence $\mathbf{X}$ to another sequence $\texttt{attention}(\mathbf{X};\theta)$

small enough

Flash attention

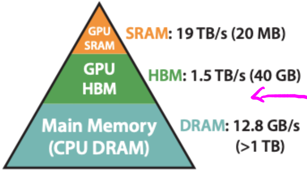
loop over block $\boxed{P_{ij}}$

$C = 1{,}000{,}000$

1: Compute $S = QK^T$

2: $P = \text{softmax}(S)$

3: $O = PV$



Memory Hierarchy with Bandwidth & Memory Size

Q K, V

Algorithm 0 Standard Attention Implementation

Require: Matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d}$ in HBM.

1: Load $\mathbf{Q}, \mathbf{K}$ by blocks from HBM, compute $\mathbf{S} = \mathbf{Q}\mathbf{K}^\top$, write $\mathbf{S}$ to HBM.

2: Read $\mathbf{S}$ from HBM, compute $\mathbf{P} = \text{softmax}(\mathbf{S})$, write $\mathbf{P}$ to HBM.

3: Load $\mathbf{P}$ and $\mathbf{V}$ by blocks from HBM, compute $\mathbf{O} = \mathbf{P}\mathbf{V}$, write $\mathbf{O}$ to HBM.

4: Return $\mathbf{O}$.

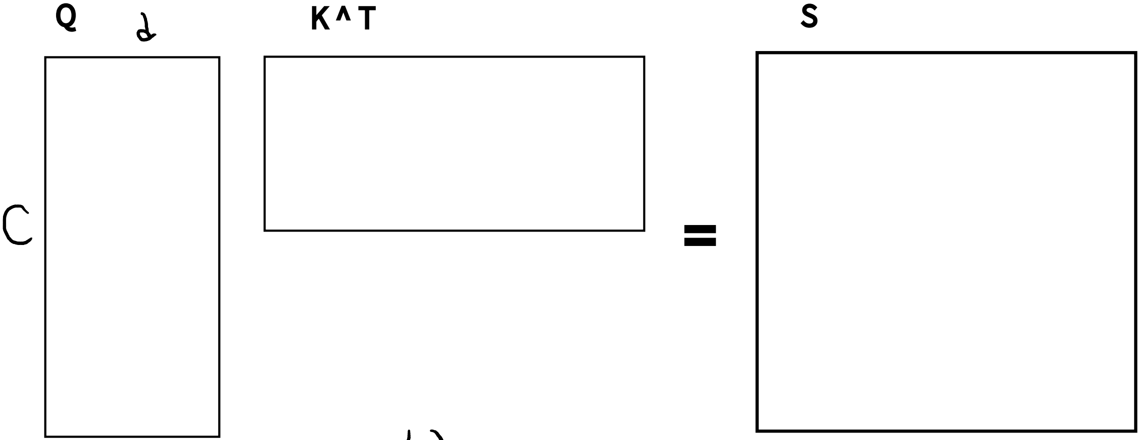
rather move Cd

Many read & write of C^2 sized matrix

From Dao et al 2022 (Flash Attention)

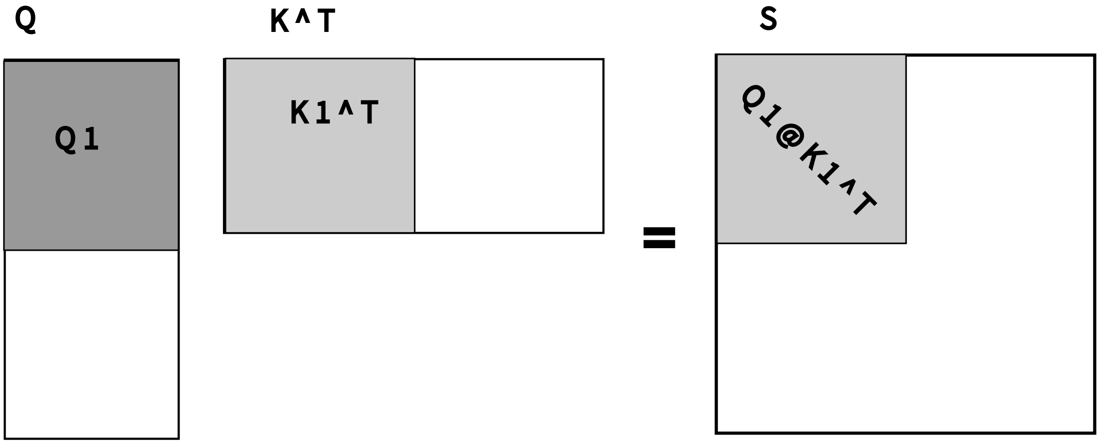
Tiling

$d \ll C$

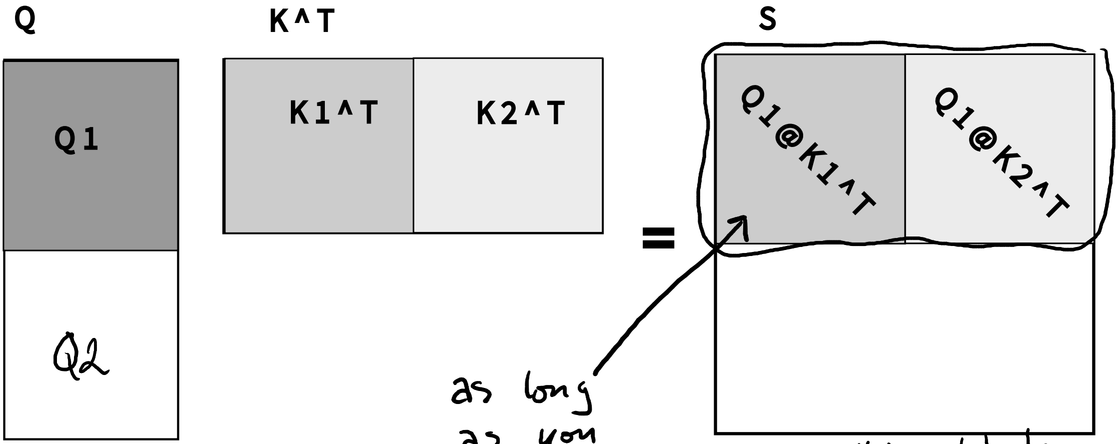


$Q = X W^{(q)}$

Tiling

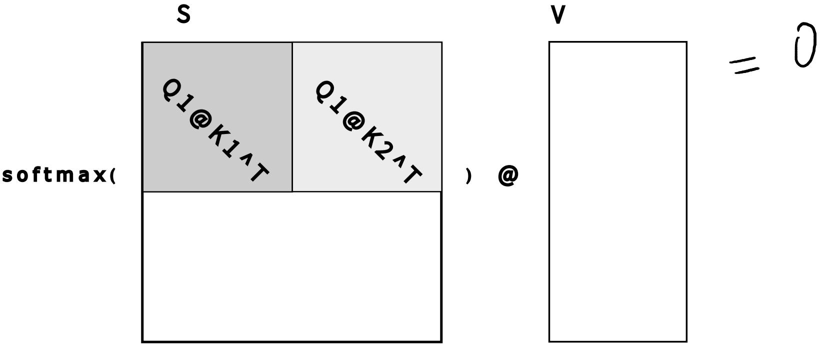


Tiling



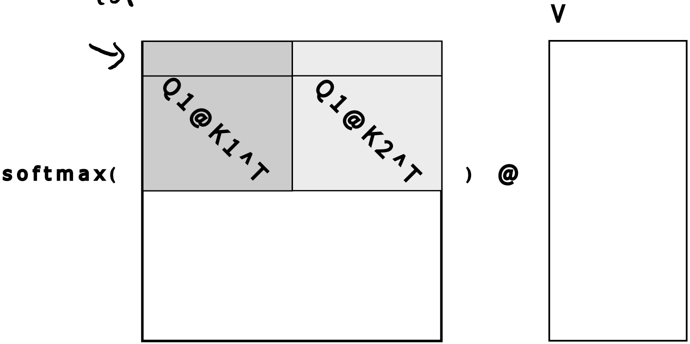
as long as you have space for this block then you can compute PV

Tiling

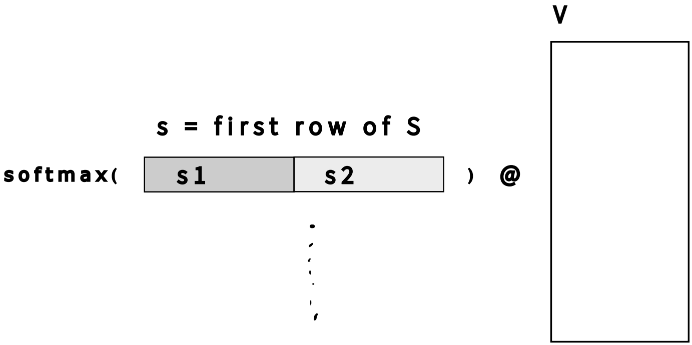


Tiling

focus on
top row



Online softmax



Online softmax

```
C = 100

s = np.random.rand(C)        = top row

I1 = np.arange(0, C//2)
I2 = np.arange(C//2, C)
s1 = s[I1]                   left
s2 = s[I2]                   right

s - np.hstack([s1,s2]) # make sure we didn't make a silly error
```

Online softmax

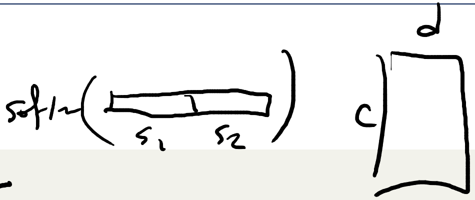
$\sum \exp(s^{(\cdot)})$

```python
def softmax(z):
    expz = np.exp(z)
    normalizer = np.sum(expz)
    return expz / normalizer, normalizer
```

probability vector

normalizer

Online softmax




```
d = 3
np.random.seed(42)
V = np.random.randn(C, d)

p, n = softmax(s)
p @ V
# >>> array([ 0.13778369, -0.16034761,  0.04310764])
```

Online softmax

softmax(s) V ✓

what if

```
# block 1
p1, n1 = softmax(s1)
O = p1 @ V[I1,:]

# block 2
p2, n2 = softmax(s2)
O = None # YOUR CODE HERE
O
```

$\text{softmax}(s_1)\ @\ \ V[I1,:] = O_1$

$\text{softmax}(s_2)\ @\ \ V[I2,:] = O_2$

How can we relate this back?

Online softmax

$$s = [\underbrace{s^1 \quad s^2}_{s1} , \underbrace{s^3 \quad s^4}_{s2}]$$

softmax

$$s1 = [s^1 \quad s^2]$$

```
# block 1
p1, n1 = softmax(s1)
O = p1 @ V[I1,:]

# block 2
p2, n2 = softmax(s2)
O = None # YOUR CODE HERE
O
```

$$\mathrm{Softmax}(s1) = \begin{bmatrix} \dfrac{\exp(s^1)}{\exp(s^1) + \exp(s^2)} \\[2ex] \dfrac{\exp(s^2)}{\exp(s^1) + \exp(s^2)} \end{bmatrix}$$

$$\frac{n_1}{n_1 + n_2} = \exp(s^1) + \dots + \exp(s^4)$$

Online softmax

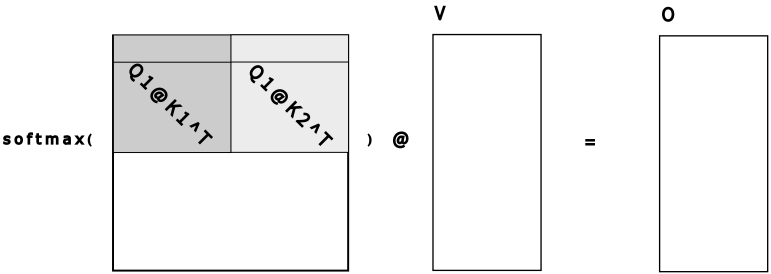
$$\frac{[\exp(s^{(1)}), \exp(s^{(2)})]}{n1}$$

$$\mathrm{softmax}(s) = \Big[\underbrace{\frac{n_1}{(n_1+n_2)} \mathrm{softmax}(s1)}_{1\text{-st half}} , \frac{n_2}{n_1+n_2} \cdot \mathrm{softmax}(s2) \Big]$$

$$\exp(s^{(1)}) + \dots + \exp(s^{(4)})$$

```
# block 1
p1, n1 = softmax(s1)
O = p1 @ V[I1,:]

# block 2
p2, n2 = softmax(s2)
O = None # YOUR CODE HERE
O
```

Summary



O_j $C \times d$
num_block

- Load $\mathbf{Q}_i$, $\mathbf{K}_j$, $\mathbf{V}_j$ one block at a time
- Compute $\mathbf{S}_{ij} = \mathbf{Q}_i \mathbf{K}_j^\top$ (S = QK usually)
- Compute $\text{softmax}(\mathbf{S}_{ij})\mathbf{V}_j$
- Update $\mathbf{O}_j$ using online softmax

small enough you can compute on chip

References I

Dao, Tri, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré (2022). "FlashAttention: Fast and memory-efficient exact attention with io-awareness". In: *Advances in Neural Information Processing Systems* 35, pp. 16344–16359.

Jain, Paras, Ajay Jain, Aniruddha Nrusimha, Amir Gholami, Pieter Abbeel, Joseph Gonzalez, Kurt Keutzer, and Ion Stoica (2020). "Checkmate: Breaking the memory wall with optimal tensor rematerialization". In: *Proceedings of Machine Learning and Systems* 2, pp. 497–511.

Sohoni, Nimit S, Christopher R Aberger, Megan Leszczynski, Jian Zhang, and Christopher Ré (2019). "Low-memory neural network training: A technical report". In: *arXiv preprint arXiv:1904.10631*.