Neural Network Preliminary
Tensor Computing

Weijie Zhao

10/12/2023

Neural Network Preliminary
Tensor Computing

- Scalar
- Vector
- Matrix
- Tensor
  - Rank
  - Dimension

Weijie Zhao

10/12/2023

Neural Network Preliminary

~~Tensor~~ Computing

Matrix

- Scalar
- Vector
- Matrix
- Tensor
  - Rank
  - Dimension

Weijie Zhao

10/12/2023

- Matrix multiplication
- Non-linear activation
- Gradient descent

- Scalar
- Vector
- Matrix
- Tensor
  - Rank
  - Dimension

Neural Network Preliminary

~~Tensor~~ Computing
Matrix

Weijie Zhao

10/12/2023

- Matrix multiplication
- Non-linear activation
- ~~Gradient descent~~ Graduate student descent

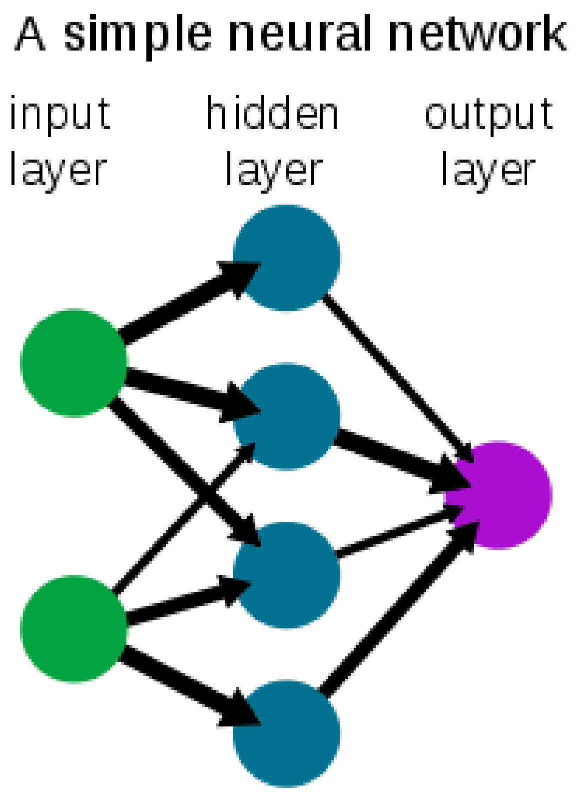
- Scalar
- Vector
- Matrix
- Tensor
  - Rank
  - Dimension

Neural Network Preliminary
~~Tensor~~ Computing
Matrix

Weijie Zhao

10/12/2023

Neural Networks



https://upload.wikimedia.org/wikipedia/commons/thumb/9/99/Neural_network_example.svg/330px-Neural_network_example.svg.png

Deep Learning Framework Implementation

- Knowing the things under the hood
- Deployment
- Deployment on emerging hardware

Deep Learning Language

- How to represent a deep neural network?

Deep Learning Language

- How to represent a deep neural network?
  - Abstraction

Deep Learning Language

- How to represent a deep neural network?
  - Abstraction
- How to implement/deploy the abstraction deep neural network?

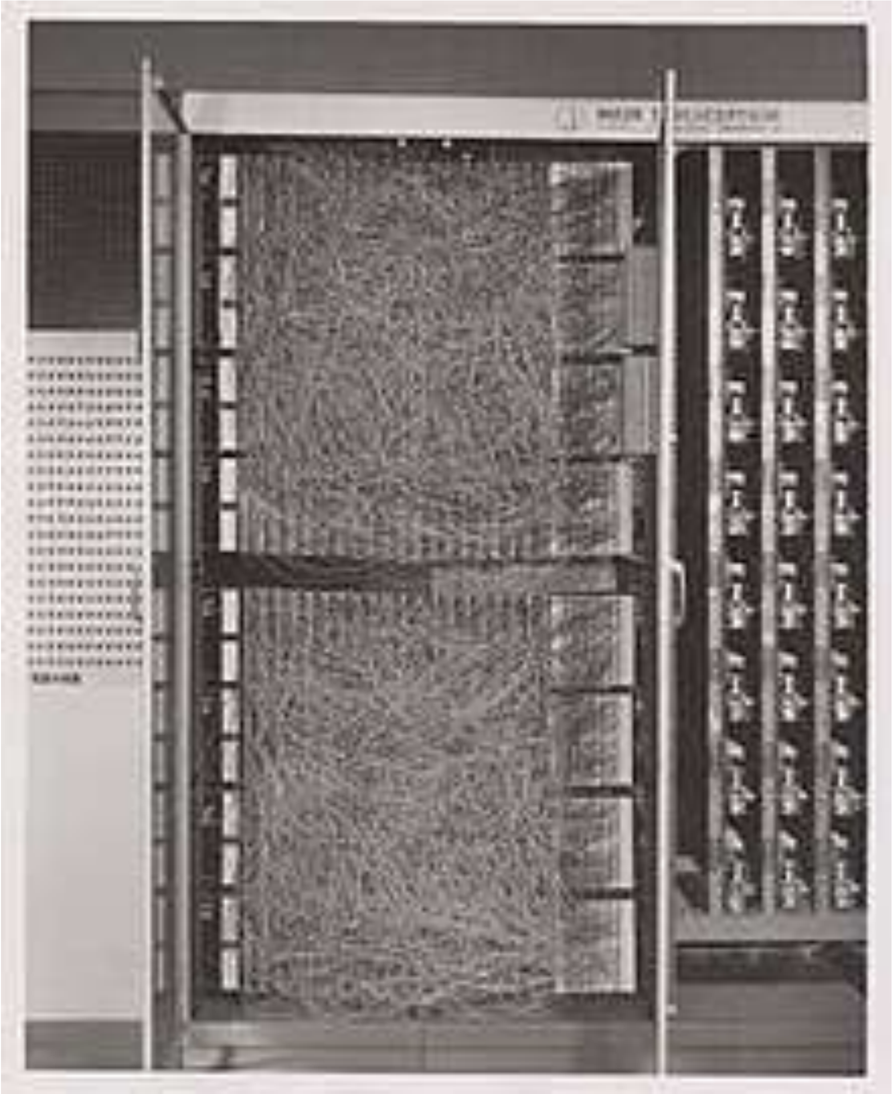
Deep Learning Language

- How to represent a deep neural network?
    - Abstraction
- How to implement/deploy the abstraction deep neural network?
    - Build tensor operations workflow
    - Implement high-performance low-level operations

Perceptron

- The perceptron was invented in 1943 by McCulloch and Pitts.
- The first implementation was a machine built in 1958 at the Cornell Aeronautical Laboratory by Frank Rosenblatt

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} + b > 0, \\ 0 & \text{otherwise} \end{cases}$$



https://upload.wikimedia.org/wikipedia/en/5/52/Mark_I_perceptron.jpeg

Perceptron

- Linear Layer
  - Matrix multiplication
  - Addition
- ReLU

Tensor Operations

- Element-wise add
- Element-wise plus
- Element-wise division
- Hadamard product
- Matrix multiplication
- Batched matrix multiplication
- More linear algebra operations…
- Collect, Scatter, Reduce…

Libraries

- Numpy
- Blas
- cuBlas
- cuSparse
- MKL
- TensorFlow
- PyTorch
- PaddlePaddle
- MXNet
- …

Lazy Evaluation and Code Generation

```
c = a + b
d = c * 2

for i = 1 to n do
        c[i] = a[i] + b[i]
for i = 1 to n do
        d[i] = c[i] * 2

for i = 1 to n do
        d[i] = (a[i] + b[i]) * 2
```

Optimizations

- Graph minimization and canonicalization
  - Constant Folding
  - Common subexpression elimination
  - Remove unnecessary operations
- Algebraic simplification and reassociation
- Copy propagation

Graph Optimizer

- Graph minimization and canonicalization
  - Constant Folding
  - Common subexpression elimination
  - Remove unnecessary operations
- Algebraic simplification and reassociation
- Copy propagation

Meta Optimizer

```
i = 0
while i < config.meta_optimizer_iterations (default=2):
    Pruning()               # Remove nodes not in fanin of outputs, unused functions
    Function()              # Function specialization & inlining, symbolic gradient inlining
    DebugStripper()*        # Remove assert, print, check_numerics
    ConstFold()             # Constant folding and materialization
    Shape()                 # Symbolic shape arithmetic
    Remapper()              # Op fusion
    Arithmetic()            # Node deduping (CSE) & arithmetic simplification
    if i==0: Layout()       # Layout optimization for GPU
    if i==0: Memory()       # Swap-out/Swap-in, Recompute*, split large nodes
    Loop()                  # Loop Invariant Node Motion*, Stack Push & Dead Node Elimination
    Dependency()            # Prune/optimize control edges, NoOp/Identity node pruning
    Custom()                # Run registered custom optimizers (e.g. TensorRT)
    i += 1
```

https://web.stanford.edu/class/cs245/slides/TFGraphOptimizationsStanford.pdf

Constant Folding Optimizer

```
do:
    InferShapesStatically()  # Fixed-point iteration with symbolic shapes
    graph_changed = MaterializeConstants()  # grad broadcast, reduction dims
    q = NodesWithKnownInputs()
    while not q.empty():
      node = q.pop()
      graph_changed |= FoldGraph(node, &q)  # Evaluate node on host
    graph_changed |= SimplifyGraph()
while graph_changed
```

Constant Folding Optimizer: SimplifyGraph()

- Removes trivial ops, e.g. identity Reshape, Transpose of 1-d tensors, Slice(x) = x, etc.
- Rewrites that enable further constant folding
- Arithmetic rewrites that rely on known shapes or inputs, e.g.
  - Constant push-down:
    - Add(c1, Add(x, c2)) => Add(x, c1 + c2)
    - ConvND(c1 * x, c2) => ConvND(x, c1 * c2)
  - Partial constfold:
    - AddN(c1, x, c2, y) => AddN(c1 + c2, x, y),
    - Concat([x, c1, c2, y]) = Concat([x, Concat([c1, c2]), y)
  - Operations with neutral & absorbing elements:
    - x * Ones(s) => Identity(x), if shape(x) == output_shape
    - x * Ones(s) => BroadcastTo(x, Shape(s)), if shape(s) == output_shape
    - Same for x + Zeros(s) , x / Ones(s), x * Zeros(s) etc.
    - Zeros(s) - y => Neg(y), if shape(y) == output_shape
    - Ones(s) / y => Recip(y) if shape(y) == output_shape

Arithmetic Optimizer

```
DedupComputations():
    do:
      stop = true
      UniqueNodes reps
      for node in graph.nodes():
        rep = reps.FindOrInsert(node, IsCommutative(node))
        if rep == node or !SafeToDedup(node, rep):
          continue
        for fanout in node.fanout():
          ReplaceInputs(fanout, node, rep)
        stop = false
    while !stop
```

Arithmetic Optimizer

- Arithmetic simplifications
  - Flattening: a+b+c+d => AddN(a, b, c, d)
  - Hoisting: AddN(x * a, b * x, x * c) => x * AddN(a+b+c)
  - Simplification to reduce number of nodes:
    - Numeric: x+x+x => 3*x
    - Logic: !(x > y) => x <= y
- Broadcast minimization
  - Example: (matrix1 + scalar1) + (matrix2 + scalar2) => (matrix1 + matrix2) + (scalar1 + scalar2)
- Better use of intrinsics
  - Matmul(Transpose(x), y) => Matmul(x, y, transpose_x=True)
- Remove redundant ops or op pairs
  - Transpose(Transpose(x, perm), inverse_perm)
  - BitCast(BitCast(x, dtype1), dtype2) => BitCast(x, dtype2)
  - Pairs of elementwise involutions f(f(x)) => x (Neg, Conj, Reciprocal, LogicalNot)
  - Repeated Idempotent ops f(f(x)) => f(x) (DeepCopy, Identity, CheckNumerics...)
- Hoist chains of unary ops at Concat/Split/SplitV
  - Concat([Exp(Cos(x)), Exp(Cos(y)), Exp(Cos(z))]) => Exp(Cos(Concat([x, y, z])))
  - [Exp(Cos(y)) for y in Split(x)] => Split(Exp(Cos(x), num_splits)

Tensor

- Dense
  - Column major
  - Row major
  - Stride
- Sparse
  - Compressed representation
  - Set intersection

Differentiation

- Numerical differentiation
- Symbolic differentiation
  - Chain rules
- Forward mode auto differentiation
- Reverse mode auto differentiation

HW2 Review

- 18/19 submissions
- 9/18 correct solutions
- Fastest: Dade Wood 124.16s
- Runner-ups:
  - Vivek Chandra Hundi Nagaraju 135.4s
  - Karamcheti Pritham 141.95s
- 7/9 correct solutions finishes in 248.32s
- X line edits in resubmission caps the score to 10-X
- The grade will be finalized by the end of 10/24

Projects 50 pts

• Cross-platform compilation	2
• High-performance implementation on CPU	5
• High-performance implementation on GPU	5
• Illegal input handling	2
• Multi-language support	1 for each language
• Non-trivial optimization/techniques	1 for each optimization
• Tasks: classification, ranking, regression, retrieval, clustering	1 for each task
• Documentation	2
• Benchmarking with baselines	5
• Proposal	10
• Demo	10
• Defend Challenging	10

Required (Proposal, Demo, Defend Challenging)

Challenging

- Each group has two chances to challenge the contribution of other group
- An incorrect challenge will cost you one chance
- You cannot challenge without any remaining chance
- Challenge is anonymous
- A successful challenge gives you half of the points you challenged

Example Projects

- Toolbox of linear classifiers with kernel method support
  - including SVM, linear regression, and logistic regression
- Gradient boosting
- Deep learning framework
- Approximate nearest neighbor search framework (KNN)

HW 3: Tensor Library

- Write a tensor library that is callable from python
- No 3rd party code is allowed. Numpy is not allowed.
- 10 test cases. Each case weights 1 pt.
- The compilation is considered failed if it does not finish in 5 minute.
- A test case is considered incorrect if it does not finish in 2 minutes.
- The numeric error of each printed value must be within 1e-3 to the correct result.
- Correct GPU solutions will get 5 pts bonus.
- The summation of the execution time across 10 cases will be used to rank correct solutions.

- Due: 10/30/2023 5:00 pm EST

Testing Environment

- ssh yourusername@granger.cs.rit.edu
- Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz
- 48 threads in total (2 sockets, 12 cores per socket, 2 threads per core)
- 251 GB memory
- GPU: Tesla P4
- pybind11 2.10.0 installed (pip3 install pybind11)

- Testing limit:
  - 8 threads taskset -c
  - 1 GPU

pybind11

```
#include <pybind11/pybind11.h>
namespace py = pybind11;
int add(int i, int j) {
    return i + j;
}
#include <pybind11/pybind11.h>

int add(int i, int j) {
    return i + j;
}

PYBIND11_MODULE(example, m) {
    m.doc() = "pybind11 example plugin"; // optional module docstring

    m.def("add", &add, "A function that adds two numbers");
}
```

```
$ python
>>> import example
>>> example.add(1, 2)
3
>>>
```

```
m.def("add", &add, "A function which adds two numbers",  py::arg("i"), py::arg("j"));
```

```
int add(int i = 1, int j = 2) {
    return i + j;
}

m.def("add", &add, "A function which adds two numbers",
      py::arg("i") = 1, py::arg("j") = 2);
```