The Image Calculator for Large-scale AI

Utku, 21 Februray 2023
Outline

- Problem definition
- Project 1 description
- Overall picture

- Image processing basics
- The Image Calculator
- Project 2 description
Outline

- Problem definition
- Project 1 description
- Overall picture
- Image processing basics
  - The Image Calculator
  - Project 2 description
Red-light cameras

New York City — 814 RL cameras

Colorectal cancer detection by tissue scans

Agricultural drones

- Black rot
- Powdery Mildew
- Apple scab
- Apple rot

[1] Image processing for smart farming: Detection of disease and fruit grading, ICIIP; 2013
1) Learn the task

Image data

Image processing algorithm
1) Learn the task

Image processing algorithm
2) Perform the task

Image data → Image processing algorithm
2) Perform the task

Image data

Image processing algorithm

Lobster!
Image processing algorithm = Convolutional Neural Networks (CNNs)

Powerful spatial-information processing machine
Efficient encoding of image data
Learning the task: Training

JPEG files ➔ RGB conversion ➔ Conv. Nnet

Host machine

GPU-accelerated machine

Conv. Nnet

RGB conversion

Host machine

GPU-accelerated machine

JPEG files

SSD

RGB conversion

CPU

GPU
Performing the task: Inference

JPEG files → RGB conversion → Conv. Nnet → GPU-accelerated machine

Lobster!
Data is bottleneck at every stage
Data is bottleneck at every stage
Data storage

JPEG files

Host machine

RGB conversion

GPU-accelerated machine

Conv. Nnet

Data is bottleneck at every stage
Outline

- Problem definition
- Project 1 description
- Overall picture
- Image processing basics
- The Image Calculator
- Project 2 description
Outline

- Problem definition
- Project 1 description
- Overall picture
- Image processing basics
- The Image Calculator
- Project 2 description

Overall picture
Every minute, >500-hours video uploaded to YouTube

[1] Figure taken from https://rb.gy/rm0hcy
Billions of parameters to learn

JPEG

JPEG is optimized for human-eye

Most images are consumed by AI models
JPEG is optimized for human-eye

Image processing algorithm

Can we have compression for AI models?
Outline

- Problem definition
- Project 1 description
- Overall picture
- Image processing basics
- The Image Calculator
- Project 2 description
Model A
Model B
Model C
Model D
Model E
Model F
What is the best compression scheme for an AI model?
Starting point: JPEG

Image

N\times N

Downsample

\frac{N}{2} \times \frac{N}{2}

Blockify

Encode

JPEG file

11100101010
10010101010
01100101011
Starting point: JPEG

JPEG is a set of design decisions
Creating a design space for compression

<table>
<thead>
<tr>
<th>Design primitives</th>
<th>Design choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling strategy:</td>
<td>2-by-2, 2-by-1, 1-by-2, 2-by-4, ...</td>
</tr>
<tr>
<td>Block size:</td>
<td>8x8, 16x16, 32x32, ...</td>
</tr>
<tr>
<td>Quantization matrix:</td>
<td>Any matrix of block size</td>
</tr>
</tbody>
</table>

Design space = all possible design choices

Each combination is a different compression scheme
Metrics

Accuracy  Time  Storage/bandwidth
Real-world models
The Image Calculator

- ML model & dataset
- Hardware & Workload
- User-defined requirements

- Design space
- World models

Optimal compression scheme
Outline

- Problem definition
- Project 1 description
- Overall picture
- Image processing basics
- The Image Calculator
- Project 2 description
Outline

Problem definition

Project 1 description

Overall picture

Image processing basics

The Image Calculator

Project 2 description
Project 1: Building accuracy model fast

- Train a few models & fit a curve
  > Slow…
Intuition 1: Each compressed version is a new dataset
Intuition 2: Images form clusters
Similar images close to each other in Euclidean space.
A class of images: a cluster
Image classification is a clustering problem.

Lobster

Hippopotamus

Ant
Examine predictive relationship between clustering-features and accuracy

C0  \rightarrow  Accuracy0
C1  \rightarrow  Accuracy1
C2  \rightarrow  Accuracy2
... 
CN  \rightarrow  AccuracyN
Examine predictive relationship between clustering-features and accuracy

C0 → Accuracy0
C1 → Accuracy1
C2 → Accuracy2
...
CN → AccuracyN

C0 → Clustering-feature0
C1 → Clustering-feature1
C2 → Clustering-feature2
...
CN → Clustering-featureN
Examine predictive relationship between clustering-features and accuracy

C0 → Accuracy0 → C0 → Clustering-feature0
C1 → Accuracy1 → C1 → Clustering-feature1
C2 → Accuracy2 → C2 → Clustering-feature2
... → ... → ...
CN → AccuracyN → CN → Clustering-featureN
Examine predictive relationship between clustering-features and accuracy
2) Create accuracy model by brute-force
   1 week

3) Explore clustering-metrics
   2 weeks

4) Examine predictive relationship
   2 weeks

5a) Move to another dataset
5b) Process-data & repeat
   6 weeks

1) Learn Pytorch
   1 week
What is success?

1) Learn Pytorch
   - 1 week

2) Create accuracy model by brute-force
   - 1 week

3) Explore clustering-metrics
   - 2 weeks

4) Examine predictive relationship
   - 2 weeks

5a) Move to another dataset
5b) Process-data & repeat
   - 6 weeks
Skills

Linux OS

GPU-accelerated computing

Storage understanding/estimation

Systems analysis/profiling

Fluent in Python & Pytorch
Outline

Problem definition

Image processing basics

The Image Calculator

Project 1 description

Project 2 description

Overall picture
Outline

- Problem definition
- Image processing basics
- The Image Calculator
- Project 1 description
- Project 2 description
- Overall picture
The Image Calculator

- ML model & dataset
- Hardware & Workload
- User-defined requirements

Optimal compression scheme
The Image Calculator

ML model & dataset → Fixed!

Hardware & Workload → The Image Calculator → Optimal compression scheme

User-defined requirements → The Image Calculator
Project 2: Building accuracy model for ML models

- Train one-by-one?
  > Way too slow…
Intuition 1: ML models have a structure
Intuition 1: ML models have a structure
Intuition 1: ML models have a structure

Conventional wisdom: larger models have larger capacity
Intuition 2: ML can share knowledge by transfer-learning
Intuition 2: ML can share knowledge by transfer learning
Intuition 2: ML can share knowledge by transfer learning
Explore predictive patterns in accuracy model

\[
\begin{align*}
&\text{M0} \rightarrow \text{Accuracy0} \\
&\text{M1} \rightarrow \text{Accuracy1} \\
&\text{M2} \rightarrow \text{Accuracy2} \\
&\ldots \\
&\text{MN} \rightarrow \text{AccuracyN}
\end{align*}
\]
Explore predictive patterns in accuracy model

M0 $\rightarrow$ Accuracy0
M1 $\rightarrow$ Accuracy1
M2 $\rightarrow$ Accuracy2
...
MN $\rightarrow$ AccuracyN

Design space of ML models

M0 M1 M2 ...
Exploit predictive patterns in accuracy model

- Intelligent sampling
- Transfer learning
1) Create accuracy model by brute-force
   - 1 week

3) Explore sampling and TL strategies
   - 3 weeks

---

1) Learn Pytorch
   - 1 week

2) Examine predictive pattern
   - 1 week

5) Move to another dataset
   - 6 weeks
What is success?

1) Create accuracy model by brute-force
   1 week

3) Explore sampling and TL strategies
   8 weeks

1) Learn Pytorch
   1 week

2) Examine predictive pattern
   1 week

5) Move to another dataset
   6 weeks
Skills

Linux OS

GPU-accelerated computing

Storage understanding/estimation

Systems analysis/profiling

Fluent in Python & Pytorch
Outline

Problem definition

Project 1 description

Overall picture

Image processing basics

The Image Calculator

Project 2 description
Outline

- Problem definition
- Project 1 description

- Image processing basics
- The Image Calculator
- Project 2 description

Overall picture
The Image Calculator

- Image dataset
- Hardware & Workload
- User-defined requirements

- Design space
- World models

- Optimal ML model
- Optimal compression scheme
Hardware & Workload
User-defined requirements
Image dataset

The Image Calculator

Design space
World models
Project 1
Optimal compression scheme
Fast world-model building for accuracy

Optimal ML model
Project 2

Optimal ML model
Optimal compression scheme

Exploring predictive accuracy models for a simplistic design space

Design space
World models

Image dataset
Hardware & Workload
User-defined requirements
JPEG files → Host machine → GPU-accelerated machine

RGB conversion

Conv. Nnet
JPEG files → RGB conversion → Conv. Nnet → GPU-accelerated machine

Data storage
Data movement

Host machine

GPU-accelerated machine

JPEG files

RGB conversion

Conv. Nnet
The Image Calculator: Data-model Co-design for Large-scale AI

- Image dataset
- Hardware & Workload
- User-defined requirements

Data design space
Data's world models
Network design space
Network's world models

Optimal ML model & compression scheme
Backup
What can we do?

- Image data

- Image processing algorithm
What can we do?

Image data

Image processing algorithm
Building accuracy model for ML models

Try each one-by-one?
> Way too slow…

Is there a predictive pattern?

Can we exploit predictive pattern?
Building accuracy model fast

Try each one-by-one?
> Way too slow...

Image classification: a clustering problem

Explore clustering-metrics vs. accuracy
Image processing algorithms need a lot of data.
Complex data

Complex algorithm

Real-world models

Storage/bandwidth

Time

Accuracy

☑
Storage/network cost

A single image ~2MB

TPC-H, lineitem table

<table>
<thead>
<tr>
<th>Orderkey</th>
<th>Partkey</th>
<th>...</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A single row ~170 bytes
Real-world models

- Storage/bandwidth
- Time
- Accuracy

Storage
SSD
WiFi
Time
Accuracy
Environmental cost

Execution time

[2] Once-for-All: Train One Network and Specialize it for Efficient Deployment, ICLR20
A novel compression scheme