Understanding and Optimizing GPU Energy Consumption of DNN Training

Jae-Won Chung
April 17th, 2023

Work done in collaboration with Jie You and Mosharaf Chowdhury
Deep Learning is Prevalent Today

- Image processing
- Speech recognition
- Machine translation
- Intelligent assistants
- Autonomous driving
- Video analytics
- Image/text generation
DNN Energy Consumption is Skyrocketing

- Re-training is commonplace (e.g. every hour)\(^2\)
- Training GPT-3 == 120 years of electricity for a household\(^1\)
- Performance optimizations oblivious of energy impact

1. U.S. EIA and Google (arXiv ’21)  2. Facebook (HPCA ’18) and Alibaba (NSDI ’22)
Existing Efforts are not Practical Enough

- New energy-efficient DNN architectures
  SqueezeNext (CVPRW ’18), ChamNet (CVPR ’19), SkyNet (MLSys ’20)

- New energy-efficient HW architectures
  TPU (ISCA ’17), EDEN (MICRO ’19), LNPU (ISSCC ’19)

- Offline profiling and power model fitting
- Confined to GPU power configuration knobs
  MPC (HPCA ’17), ODPP (CCGRID ’20), GPOEO (TPDS ’22)
Understanding GPU Energy Consumption

**Energy to Accuracy (ETA)**
- Energy needed to reach the user-specified target accuracy
- Energy-counterpart of *Time to Accuracy* (TTA)
Understanding GPU Energy Consumption

\[ \text{ETA} = \text{TTA} \times \text{AvgPower} \]

- \text{ETA}: Joule
- \text{TTA}: Second
- \text{AvgPower}: Watt
Understanding GPU Energy Consumption

\[ \text{ETA} = \#\text{Epochs} \times \text{EpochTime} \times \text{AvgPower} \]

- **#Epochs**: Job side
- **EpochTime**: Job side
- **AvgPower**: GPU side
- **ETA**: Joule
- **Joule**: Watt
- **Watt**: Second

Motivation | 7
Opportunity for Energy Savings

Sweep of feasible batch sizes and power limits

- Default
- Best batch size
- Best power limit
- Best of both

Normalized Energy to Accuracy (ETA)

24 ~ 75% energy reduction

Measured on an NVIDIA V100 GPU. Training terminates when the DNN reaches its original target accuracy.
Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100 GPU. Similar trends found across 6 DL workloads and 4 GPU generations.
Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100 GPU. Similar trends found across 6 DL workloads and 4 GPU generations.
Relationship Between Time and Energy

- Feasible ▲ Default ▶ Pareto Front

1. Time and energy minimized by different knobs
2. Efficient time and energy show a trade-off

Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100 GPU. Similar trends found across 6 DL workloads and 4 GPU generations.

Motivation | 11
Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100 GPU. Similar trends found across 6 DL workloads and 4 GPU generations.

Motivation | 12

Cost = $\eta \cdot \text{ETA} + (1 - \eta) \cdot \text{MaxPower} \cdot \text{TTA}$
Relationship Between Time and Energy

![Diagram showing the relationship between time and energy with different points and lines indicating feasible, default, and Pareto Front areas.]

Which yellow point is the best?

\[ \text{Cost} = \eta \cdot \text{ETA} + (1 - \eta) \cdot \text{MaxPower} \cdot \text{TTA} \]

Results from training DeepSpeech2 on LibriSpeech on an NVIDIA V100 GPU. Similar trends found across 6 DL workloads and 4 GPU generations.
An Energy Optimization Framework for DNN Training

Optimizes the cost

- of an arbitrary DNN model
- on an arbitrary GPU type
- in an efficient manner

without any

- offline profiling,
- hardware modification, or
- accuracy degradation
Overall Workflow

Re-training jobs are opportunity for exploration!
Overall Workflow

Re-training jobs are opportunity for exploration!

1. Decoupling Variables
2. Power Limit Optimizer
3. Batch Size Optimizer
I. Decoupling Batch Size and Power Limit

\[ \text{ETA} = \#\text{Epochs} \times \text{EpochTime} \times \text{AvgPower} \]

- **ETA**: Joule
- **\#Epochs**: random
- **EpochTime**: Second
- **AvgPower**: Watt

**Batch Size**
- Deterministic
- Batch Size Optimizer

**Power Limit**
- Online
- Power Limit Optimizer
2. Power Limit Optimizer

Just-in-time online profiler
- Profiles the power and throughput of each power limit
- Five seconds per power limit is enough

Low overhead
- Profile only once for each batch size
- Profiling contributes to the training process
3. Batch Size Optimizer

A good solution must
1. incorporate the stochasticity of DNN training, and
2. intelligently trade-off exploration and exploitation

Multi-Armed Bandit
1. Models cost as a Gaussian random variable
2. Automatically controls exploration and exploitation

Cost = \eta \cdot ETA + (1 - \eta) \cdot MaxPower \cdot TTA
# Workloads and GPU Generations

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset</th>
<th>DNN</th>
<th>GPU</th>
<th>Arch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Recognition</td>
<td>LibriSpeech</td>
<td>DeepSpeech2</td>
<td>NVIDIA A40</td>
<td>Ampere</td>
</tr>
<tr>
<td>Question Answering</td>
<td>SQuAD</td>
<td>BERT</td>
<td>NVIDIA V100</td>
<td>Volta</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>Sentiment140</td>
<td>BERT</td>
<td>NVIDIA RTX6000</td>
<td>Turing</td>
</tr>
<tr>
<td>Image Classification</td>
<td>ImageNet</td>
<td>ResNet-50</td>
<td>NVIDIA P100</td>
<td>Pascal</td>
</tr>
<tr>
<td>Image Classification</td>
<td>CIFAR-100</td>
<td>ShuffleNet-v2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recommendation</td>
<td>MovieLens-1M</td>
<td>NeuMF</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Zeus in Action

Grid Search

Converging Point

Power Limit (W)

Batch Size

250 - 225 - 200 - 175 - 150 - 125 - 100 -
8 12 16 24 32 48 56 64 72 96 128 156 192 250

DeepSpeech2 trained on LibriSpeech on an NVIDIA V100 GPU.

Zeus

Converging Point

Power Limit (W)

Batch Size

250 - 225 - 200 - 175 - 150 - 125 - 100 -
8 12 16 24 32 48 56 64 72 96 128 156 192 250

Search Path Training Cost (darker means better)
Zeus Leads to Large Benefits

15 ~ 76% energy reduction
Up to 60% time reduction

Results obtained on an NVIDIA V100 GPU
Demo: Stable Diffusion

https://youtu.be/MzlF5XNRSJY
Conclusion

DNN

- Works on arbitrary DNN models

GPU

- Works without modifying existing hardware

Energy

- Fully online with JIT profiling and MAB
- Jointly optimizes both job- and GPU-side configurations
https://ml.energy/zeus