**ooc_cuDNN** : A Deep Learning Library Supporting CNNs over GPU Memory Capacity

Yuki Ito, Ryo Matsumiya, Toshibo Endo (Tokyo Institute of Technology)

---

**Background**

- Convolutional neural networks (CNNs) are used in many fields.
  - Image recognition, Image processing, speech recognition, etc...

**cuDNN** [1] library can accelerate computation of CNNs
- Developed by NVIDIA
- Used by many deep learning frameworks
- Use graphic processing units (GPUs) effectively

**Motivation**
- It is hard for large scale CNNs to be computed using cuDNN
- cuDNN can use GPU memory only
- GPU memory capacity is limited
  - Even computation of one layer may run out of GPU memory

---

**Our solution**

- We designed and implemented ooc_cuDNN [2] library.
- ooc_cuDNN (out-of-core cuDNN) supports large scale CNNs
  - Compatible with cuDNN
  - Enable to compute CNNs that exceed GPU memory capacity
  - Use both GPU and CPU memory
  - Divide layers and filters
  - Each layer (or filter) is put on GPU or CPU memory
  - Swap data between CPU and GPU memory
  - Overlap CPU-GPU communication and computation

---

**Optimization(1) : Auto-tuning division sizes**

- Performance of ooc_cuDNN is affected by each division size.
  - Make performance model
  - Optimize division size based on the model.

---

**Optimization(2) : Fusion of computations**

- Performance of low complexity computations is too low in ooc_cuDNN.
  - In those computations, communication can not be hidden completely.
  - Provide fused functions that perform high complexity computations and low complexity computations at once.

---

**Evaluation**

- Apply ooc_cuDNN to CNN application
  - Forward and Backward of VGG16[3]
  - The required memory size increases according to batch size.
- Experiment with Tesla P100
  - ooc_cuDNN enables to compute CNN exceeding GPU memory capacity.

---

**Integrating with deep learning framework**

- We implemented ooc_Caffe2 (Caffe2 with ooc_cuDNN).
  - Caffe2[4] is a deep learning framework developed by Facebook.
  - Not support ooc_cuDNN’s fused functions in current design.
- For comparison, we implemented unified_Caffe2.
  - Use original cuDNN, and allocate data as Unified Memory.
  - Unified memory supports data exceeding GPU memory capacity by swapping mechanism between CPU and GPU.
  - ooc_Caffe2 is > x1.7 faster in out-of-core cases.

---

**Future work**

- Optimization considering the entire CNN
  - Which data should be put on CPU memory?
  - Which computation should be fused?
- Improve ooc_Caffe2
  - Use fused functions
  - Support distributed computation

---


This work is supported by JST-CREST, “Software Technology that Deals with Deeper Memory Hierarchy in Post-petascale Era”