# Kazuki Osawa

## CONTACT INFORMATION

2-12-1 i7-2 103 Ookayama, Meguro-ku, Tokyo 152-8550, JAPAN
Email: oosawa.k.ad@m.titech.ac.jp

## EDUCATION

March 2021 Tokyo Institute of Technology (Japan), Ph.D. (expected) in Computer Science
March 2018 Tokyo Institute of Technology (Japan), M.S. in Computer Science
March 2016 Tokyo Institute of Technology (Japan), B.S. in Computer Science

## RESEARCH INTERESTS

- Second-order optimization methods for deep learning
- Large-scale distributed computing
- Bayesian deep learning
- Low-rank approximation

## RESEARCH EXPERIENCE

A*STAR Research Attachment Program (ARAP), A*STAR, Oct 2018 – present

*Advisor:* Chuan-Sheng Foo and Vijay Chandrasekhar
*Area:* Machine Learning

- Improving second-order optimization methods of deep learning from the perspective of the loss landscape

Ph.D.'s Research, Tokyo Institute of Technology, Apr 2018 – present

*Advisor:* Rio Yokota
*Area:* Machine Learning & High-Performance Computing

- Implemented K-FAC (Kronecker-Factored Approximation of the Fisher Information Matrix) on Chainer (a framework developed by Preferred Networks, Japan)
- Implemented distributed K-FAC on ChainerMN (multi-node) for extremely large min-batch (128K) training of CNN for ImageNet (1K class) dataset
- Distributed second-order optimization on the supercomputer with over 4,000GPUs
- Analyzing the effectiveness of second-order optimization for Variational Inference for learning of CNNs

Master's Research, Tokyo Institute of Technology, Apr 2016 – Mar 2018

*Advisor:* Rio Yokota
*Area:* Machine Learning & High-Performance Computing

- Applied Low-Rank Approximation of the weight tensors of Convolutional Neural Networks to accelerate inference
- Analyzed the feature of the tensors (e.g. Singular Values) and evaluated trade-off

between the speed-up and the accuracy of image recognition
- Analyzed the effectiveness of the second-order optimization methods such as K-FAC to achieve fast computing while approximating exact Natural Gradient for Deep Neural Networks

R&D Internship, Nefrock Lab Ookayama, Nefrock Inc. (Japan), Jun 2016 – Dec 2016
- Developed a Convolutional Neural Network for the application to judge compatibility between eyeglasses and human face for Japanese eyeglasses manufacturer (https://brain.jins.com/)
- Fine-tuned the existing model of CNN to fit to our application using evaluation data made by the staff of the manufacturer

R&D Internship, Central Research Laboratories, NEC (Japan), Aug 2016 – Oct 2016
- Analyzed and modified computing algorithm in existing Deep Learning framework so that we can utilize parallel distributed processing infrastructure
- Translated CUDA code to C++ code to utilize vector type computing machine

Undergraduate Research, Tokyo Institute of Technology, Apr 2015 – Mar 2016

*Advisor:* Isao Yamada
*Area:* Optimization & Signal Processing

- Developed algorithm of Spline Smoothing to achieve minimization of Total Variation
- Realized a solution of function data analysis with adopting Total Variation as smoothness criterion
- Derived the Proximity Operator of Total Variation of piecewise polynomial (spline function) so that we can apply it to ADMM

Research Practice Program, Tokyo Institute of Technology, Aug 2014 – Oct 2014

*Advisor:* Koichi Shinoda
*Area:* Pattern Recognition

- Developed finger vein authentication system
- Implemented and analyzed the existing algorithm

## MEMBERSHIPS

Association for Computing Machinery (ACM)
Society for Industrial and Applied Mathematics (SIAM)
Information Processing Society of Japan (IPSJ)
The Japan Society for Computational Engineering and Science (JSCES)

## TEACHING EXPERIENCE

Teaching Assistant, NVIDIA Deep Learning Institute Day 2017, Tokyo, Apr 2017
- Teach participants how to use NVIDIA DIGITS and how to build CNN and RNN
- Teach basic Python/Jupyter techniques

Teaching Assistant, SuperCon2016 at Tokyo Institute of Technology, Aug 2016

- Teach high school students how to use Linux system and how to run scripts on Supercomputer TSUBAME 2.5 at Tokyo Institute of Technology
- Teach basic C++/CUDA C techniques for Supercomputing Contest

## SCHOLARSHIPS & GRANT-IN-AIDS

The Nakajima Foundation, Ph.D. Scholarship (2018-2023, tuition and living expenses)
Japan Student Services Organization, Master's Scholarship (2016-2018, 2,112,000 yen)
International Information Science Foundation, Overseas Dispatch of Researchers (2017, 180,000 yen)
SC17, Student Volunteers (2017, $1,500 USD)

## PUBLICATIONS

### International (refereed)

1. **Kazuki Osawa** and Rio Yokota. "Evaluating the Compression Efficiency of the Filters in Convolutional Neural Networks", *Artificial Neural Networks and Machine Learning – ICANN 2017*, pp 459-466, Springer 2017.
2. **Kazuki Osawa**, Akira Sekiya, Hiroki Naganuma, Rio Yokota. "Accelerating Matrix Multiplication in Deep Learning by Using Low-Rank Approximation", *2017 International Conference on High Performance Computing & Simulation (HPCS)*, pp 186-192, IEEE 2017.

### Domestic (refereed)

3. **Kazuki Osawa**, Akira Sekiya, Hiroki Naganuma, Rio Yokota. "Accelerating Convolutional Neural Networks Using Low-Rank Tensor Decomposition", *Pattern Recognition and Media Understanding*, Kumamoto, Japan, Oct. 2017.
4. Hiroki Naganuma, Akira Sekiya, **Kazuki Osawa**, Hiroyuki Ootomo, Yuji Kuwamura, Rio Yokota. "Improvement of speed using low precision arithmetic in deep learning and performance evaluation of accelerator", *Pattern Recognition and Media Understanding*, Oct. 2017.
5. Hiroki Naganuma, **Kazuki Osawa**, Akira Sekiya, Rio Yokota. "Acceleration of Compressed Models in Deep Learning Using Half Precision Arithmetic", *Japan Society for Industrial and Applied Mathematics Annual Meeting*, Sep. 2017.
6. **Kazuki Osawa**, Akira Sekiya, Hiroki Naganuma, Rio Yokota. "Accelerating Convolutional Neural Networks using Low-Rank Approximation", *The 22nd Conference on The Japan Society for Computational Engineering and Science*, Jun. 2017.

### Domestic (non-refereed)

7. Sekiya Akira, **Kazuki Osawa**, Hiroki Naganuma, Rio Yokota. "Accelerating Matrix Multiplication for Deep Learning Using Low-Rank Approximation", *158h Research Presentation Seminar in High-Performance Computing*, Mar. 2017.