YUNSEONG LEE

Gwanak-ro1, Gwanak-gu, Seoul, South Korea
✉ *yunseong@snu.ac.kr*

Education

2014–2020 **PhD**, *Seoul National University*, Seoul, South Korea.
Computer Science and Engineering (advisor: Byung-Gon Chun)

2007–2014 **Bachelor**, *Seoul National University*, Seoul, South Korea.
Computer Science and Engineering

Work Experience

2014–present **Research Assistant**, *Software Platform Lab, SNU*, Seoul, South Korea.
I have studied and built large-scale data processing systems, recently focusing on building Machine Learning (ML) and Deep Learning (DL) inference systems. I have participated in the following projects:

- **Pretzel**: A white box ML inference system for model pipelines. Unlike previous black box systems, Pretzel exploits knowledge of models obtained from the training phase. This knowledge enables Pretzel to optimize the end-to-end execution plan of models. Pretzel also considers multiple models to be served together and schedules CPU/memory resources more efficiently.
- **RnB (Replicate-and-Batch)**: A multi-GPU DL inference system for DNN pipelines. DNN pipelines often consist of sub-networks with heterogeneous performance characteristics (e.g., latency). To optimize the throughput and latency, RnB replicates each sub-network to a different number of GPUs and batches (or splits) inputs of each sub-network based on their sizes. RnB is built on PyTorch.
- **Splash**: A DL system for collaborative inference in a distributed environment with embedded devices and a *hub* machine. Splash aims to run highly accurate (but heavy) DNN models on computationally weak devices with assistance of the strong hub. Splash applies compression techniques such as Tucker decomposition to reduce the inference latency. Splash currently supports TensorFlow and Caffe2.
- **ICCV 2019 AIA (AI Acceleration) Challenge**: I participated in the AIA challenge, specifically in the DSP (Digital Signal Processor) track. I ran a DNN model on TFLite (with Android NNAPI) and accelerated the model by utilizing DSPs (e.g., NPU) with optimizations such as quantization.

2014–present **Project Management Committee (PMC)**, *Apache REEF*, Apache Software Foundation.
I am one of the PMC members of Apache REEF. I worked as the release manager of REEF v0.14.

2017/6–2017/9 **Research Intern**, *Microsoft*, Redmond, US.
I investigated how to optimize ML inference systems with white box approaches. The research was published in ICCD, NIPS ML Systems workshop, SysML, OSDI, and IEEE Data Engineering bulletin.

2015/9–2016/2 **Research Intern**, *Microsoft Research Asia*, Beijing, China.
I participated in the Pado project, a data processing system for handling transient resources in data centers. This research was published in EuroSys.

Publications

[1] W.-Y. Lee, **Y. Lee**, J. S. Jeong, G.-I. Yu, J. Y. Kim, H. J. Park, B. Jeon, W. Song, G. Kim, M. Weimer, B. Cho, and B.-G. Chun. *Automating System Configuration of Distributed Machine Learning*. ICDCS, 2019.

[2] **Y. Lee**, A. Scolari, B.-G. Chun, M. Weimer, and M. Interlandi. *From the Edge to the Cloud: Model Serving in ML.NET*. IEEE Data Engineering Bulletin, december edition, 2018.

[3] **Y. Lee**, A. Scolari, B.-G. Chun, M. D. Santambrogio, M. Weimer, and M. Interlandi. *Opening the Black Box of Machine Learning Prediction Serving Systems*. OSDI, 2018.

[4] **Y. Lee**, A. Scolari, M. Interlandi, M. Weimer, and B.-G. Chun. *Towards High-Performance Prediction Serving Systems*. SysML, 2018.

[5] **Y. Lee**, A. Scolari, M. Interlandi, M. Weimer, and B.-G. Chun. *Towards High-Performance Prediction Serving Systems*. NIPS ML Systems Workshop, 2017.

[6] A. Scolari, **Y. Lee**, M. Weimer, and M. Interlandi. *Towards Accelerating Generic Machine Learning Prediction Pipelines*. ICCD, 2017.

[7] Y. Yang, G.-W. Kim, W. W. Song, **Y. Lee**, A. Chung, Z. Qian, B. Cho, and B.-G. Chun. *Pado: A Data Processing Engine for Harnessing Transient Resources in Datacenters*. EuroSys, 2017.

[8] B.-G. Chun, T. Condie, Y. Chen, B. Cho, A. Chung, C. Curino, C. Douglas, M. Interlandi, B. Jeon, J. S. Jeong, G.-W. Lee, **Y. Lee**, T. Majestro, D. Malkhi, S. Matusevych, B. Myers, M. Mykhailova, S. Narayanamurthy, J. Noor, R. Ramakrishnan, S. Rao, R. Sears, B. Sezgin, T.-G. Um, J. Wang, M. Weimer, and Y. Yang. *Apache REEF: Retainable Evaluator Execution Framework*. ACM Transactions of Computer Systems, october edition, 2017.

[9] B.-G. Chun, B. Cho, B. Jeon, J. S. Jeong, G. Kim, J. Y. Kim, W.-Y. Lee, **Y. S. Lee**, M. Weimer, and G.-I. Yu. *Dolphin: Runtime Optimization for Distributed Machine Learning*. ICML ML Sys. Workshop, 2016.

[10] M. Weimer, Y. Chen, B.-G. Chun, T. Condie, C. Curino, C. Douglas, **Y. Lee**, T. Majestro, D. Malkhi, S. Matusevych, B. Myers, S. Narayanamurthy, R. Ramakrishnan, S. Rao, R. Sears, B. Sezgin, and J. Wang. *REEF: Retainable Evaluator Execution Framework*. SIGMOD, 2015.

[11] J. S. Jeong, W.-Y. Lee, **Y. Lee**, Y. Yang, B. Cho, and B.-G. Chun. *Elastic Memory: Bring Elasticity Back To In-Memory Big Data Analytics*. HotOS, 2015.