I am a second-year Computer Science Master student at Georgia Tech under supervision of Professor Hyesoon Kim and Joy Arulraj. I want to pursue a Ph.D. in Computer Science. I aspire to become a professor to build efficient systems. My research interests include computer system, computer architecture, and machine learning. I have contributed to publications in conferences, such as IROS'2018, SysML'2019, and IISWC'2019. I also have two works at SIGMOD'2020 and CVPR'2020, which are currently under review.

My motivation for doing research stems from my summer project with Professor Hyesoon Kim. In that project, I prototyped a distributed system on ten edge devices. The system provides service to a deep neural network (DNN) based visual data analytics applications. However, the system performance is inefficient because the algorithm of the application is compute intensive and the edge devices themselves are resource constrained (on both memory space and compute power). To improve the system performance, we proposed to distributively execute the computation on multiple devices. I mainly worked on modifying the high level algorithm to fit to distributed computation and implementing system primitives (data transmission through RPC and data synchronization, etc). We published the work at IROS'2018 [1]. I extended the system to object detection application and showcased the system at SysML'2019 [5]. From this project, I learned that a system can be constrained by both high level algorithms and low level hardware resources. As a system researcher, I strove to better understand the low level hardware primitives, and design system tailored high level algorithm, and combine those techniques together to build a more efficient system.

To better understand the hardware constraints on application, I studied the performance of DNNs on different hardware platforms. We realized that the Computer Science community lacks a systematic characterization of the DNNs on different platforms, especially the edge devices. In addition to that, a mismatch between our profiled performance and reported performance from vendors sometimes existed. We decided to conduct a comprehensive performance analysis for available platforms. In this work, I led the engineering efforts on integrating and analyzing popular DNNs with different devices. I conducted timing, memory usage, and power analysises for different workloads, in which I extensively used cProfile to identify the bottleneck in each case. We published this work at IISWC'2019 [2] as the best paper nominees. This work solidified my view that the co-design of high level algorithm and low level system primitives delivers the most optimal performance, so the application design should always incorporate the characteristics of low level primitives.

Motivated by that, I attempted to optimize the DNN workload from a system researcher's perspective. First of all, DNN design is usually tailored for accuracy. Secondly, the typical single-chain DNN architecture has dependencies which is sub-optimal for deployment on multi-core or distributed systems. To bridge the gap, I endeavered to propose a new DNN design methodology, which optimizes performance based on hardware platforms. In this work, we used randomly connected graphs to construct DNNs. We formulated the problem as graph partition problem, in which we estimate the resulted performance and communication cost from a particular partition. The new architecture achieves lower inter-processes communication and lower latency by searching the optimal partition of the randomly connected graph. This work is currently under review at CVPR'2020 [4].

After learn both algorithm designs and hardware concepts, I aspired to build a sophisticated system with learned ideas. I realized that most of the visual data analytics system optimizations were done for a cloud-based system, so I attempted to study accelerating the system on edge devices. With the help from Professor Joy Arulraj, I designed an edge centric visual data analytics system from scratch. Two challenges of building such a system are the limited computation resource on edge and limited bandwidth from edge to cloud. To strike the first challenge, I designed a specialized visual data analytics DNN to leverage computation cost. To leverage communication cost, I built a customized compression scheme on top of CSR format. Moreover, a lossy compression scheme is added to reduce the data transfer overhead further. The third component that I proposed is an edge tailored scheduler. It dynamically searches for the optimal workload placement choice and schedules the workload efficiently on appropriate device. We demonstrated that edge resource can be efficiently utilized with the careful co-design of high level algorithm and low level system implementation. This work is currently under review at SIGMOD'2020 [3].

These research experiences have shaped me into a proficient system researcher. While I am open to any system problem, I am particularly interested in building systems for data intensive applications. I have two particular directions (*bottom-up* and *top-down*) in mind that I strive to optimize. I want to work on building special systems either on edge devices or cloud for data intensive applications (*bottom-up*). I also want to reshape the algorithms/applications to achieve better fit to low level systems (*top-down*).

I want to continue my study at UC Berkeley. Though I am open to a variety of research, there are several professors at UC Berkeley whose works are especially appealing to me. Because I am currently working on multi-camera system optimizations, I am very interested in **Professor Joseph Gonzalez**'s paper *"Scaling Video Analytics Systems to Large Camera Deployments"*. I also really enjoy reading **Professor Ion Stoica**'s paper *"Chameleon: Video Analytics at Scale via Adaptive Configurations and Cross-Camera Correlations"* It would be really exciting to study at EECS department at UC Berkeley. I look forward to becoming a mature system researcher after my graduate study.

References

[1] Ramyad Hadidi, **Jiashen Cao**, Matthew Woodward, Michael Ryoo, and Hyesoon Kim. Distributed perception by collaborative robots. In *Proceedings of IROS*, 2018.

[2] Ramyad Hadidi, **Jiashen Cao**, Yilun Xie, Bahar Asgari, Tushar Krishna, and Hyesoon Kim. Characterizing the deployment of deep neural networks on commercial edge devices. In *Proceedings of IISWC*, 2019.

[3] **Jiashen Cao**, Ramyad Hadidi, Joy Arulraj, and Hyesoon Kim. Accelerating visual data analytics using edge devices. In *Submissions of SIGMOD*, 2020.

[4] **Jiashen Cao***, Ramyad Hadidi*, Michael Ryoo, and Hyesoon Kim. Parallelnets: Increasing inference performance with parallel architectures for image recognition. In *Submissions of CVPR*, 2020.

[5] **Jiashen Cao**, Fei Wu, Ramyad Hadidi, Lixing Liu, Tushar Krishna, Michael Ryoo, and Hyesoon Kim. An edge-centric scalable intelligent framework to collaboratively execute dnn. In *Proceedings of Demo - SysML*, 2019.