Systems and Architecture Support for Large-Scale Visual Search

I propose to design a scalable system to catalog and search the multimedia data of the Internet by rethinking the roles of the memory subsystem and networking layer.

(1) Broader Impact. As the Internet shifts towards unprecedented multimedia content creation and consumption, significant innovation at all levels of the computing stack is imperative. In 2013 alone, approximately 127 billion images and 620 million videos were uploaded to Facebook and YouTube, respectively [1,2]. In the same year, traffic volumes from Netflix and YouTube accounted for more than 50% of global Internet traffic [3]. This explosive growth in multimedia content has created a landscape where **visual search engines will become essential to reason about the volume of data available to the end user**. To date, visual search engines exist only as research prototypes in part because techniques in text-based search do not apply to the unstructured nature of multimedia data. The problem is further exacerbated by the larger I/O bandwidth incurred by multimedia content. To enable visual search at large-scale, the following key challenges must be addressed.

Feature Storage and Retrieval. Fast feature storage and retrieval is critical to satisfy the aggressive demands of low-latency in user-facing applications. I propose to intelligently: (1) exploit locality by sharding similar feature vectors and (2) minimize data movement across the networking layer and within the memory subsystem.

Indexing. The index is a data structure that incurs an expensive one-time cost that can be amortized with many searches. However, the time it takes to build an index grows linearly to the cardinality and dimensionality of the data. Furthermore, the index must adapt and be rebuilt when new visual content is uploaded to the database. I propose to: (1) design a scalable, distributed data structure for visual search, (2) develop a methodology for incremental reindexing, and (3) integrate application awareness into the systems architecture to identify opportunities for further optimization.

To address the aforementioned challenges, I will draw from the following approaches: (1) introduce architectural support to the application via **in-memory** computation, (2) **co-design** hardware and software subsystems to enable incremental reindexing, and (3) evaluate a lightweight **accelerator** design for k-nearest neighbors.

(2) Background. A visual search engine uses the content of images or videos as search input and returns media of similar visual characteristics. Visual search consists of two phases: feature extraction and similarity search. In feature extraction, a convolutional neural network transforms the content of a media object and generates a feature vector (e.g., 4096 features for images and 131,072 features for 5-minute videos [4, 5]). Similarity search then compares each feature pairwise using the k-nearest neighbor (kNN) algorithm. I focus exclusively on synergistic co-design of hardware and software to support the kNN algorithm.

Prototype for Visual Search. The prototype will leverage existing state-of-the-art library implementations of k-nearest neighbors as a baseline. I will push the performance of the system purely from a software standpoint using traditional CPU optimization techniques (i.e., locality, cache blocking, prefetching). Then, I will extend the system to the clusterlevel by distributing the application workload across many systems.

Characterization, Analysis and Optimization of kNN. I will characterize the Fast Library for Approximate Nearest Neighbors (FLANN) library for its three core kNN algorithms: *parallel randomized kd-trees*, *priority search k-means*, and *linear search*. I will analyze the trade-offs of each algorithm and evaluate system performance based on latency and throughput targets. Round-trip latency will be critical to meet an acceptable service-level objective. Throughput analysis will bound the number of concurrent requests for a single node.

Lightweight FPGA-based Accelerator for kNN. The kNN algorithm performs pairwise Euclidean distance calculations between feature vectors. Due to its low computational intensity, this computation is highly memory-bound. I propose architectural extensions in the form of floating point adders and multiplication units embedded within the main DRAM memory to minimize data movement and push the distance calculation closer to memory. I will evaluate this approach compared to loading and computing feature vectors in on-chip memory.

Scalable Indexing Structures for Visual Features. The lack of a scalable indexing feature in FLANN significantly limits performance in distributed systems. First, I will identify opportunities for scalable indexing and implement the index at the networking layer. I will co-design the software indexing structure with physical hardware routers to accelerate search time. As new multimedia content is uploaded, the search index must also accommodate and adapt to make this new content searchable. I will add extensions to the current FLANN library to support incremental indexing.

(3) Execution Plan. This is a multi-year project jointly with my advisors, Luis Ceze and Mark Oskin from Computer Architecture and Ali Farhadi from Computer Vision. In the span of three years, I will: (1) design and implement an end-to-end visual search system for images and videos that exceeds single-node DRAM capacity and saturates network throughput, (2) evaluate the tradeoffs between different approaches for kNN (i.e., randomized kd-tree, k-means, linear) for high cardinality and high dimensionality feature data, (3) evaluate the efficacy of a lightweight accelerator for kNN via cycle-accurate simulation, and (4) develop a scalable, distributed indexing structure with extensions for incremental indexing.

References

[1] Internet.Org, "A Focus on Efficiency: A Whitepaper from Facebook, Ericsson and Qualcomm," 2013. [Online]. Available: http://internet.org/efficiencypaper

[2] YouTube, "Statistics - YouTube," 2014. [Online]. Available: https://www.youtube.com/yt/press/statistics.html

[3] Intel, "What Happens in an Internet Minute?" Oct. 2014. [Online]. Available: http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html

[4] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[5] H. Wang *et al.*, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 3169–3176.

(1) Personal Statement

Computer Science is More Than Just Writing Code. I was fortunate to work with a leading technology company on one of its simulation projects during an internship following my freshman year. The company flew me to a naval base in New Jersey to deploy our team's product to the U.S. government. It was spectacular. I experienced a large-scale visual simulation of an ocean complete with submarines, boats, and tugs projected onto a dome-shaped room with a full 360-degree view. It modeled reality so well that I could not distinguish between the authentic and the artificial. Those beautifully rendered images was a product of the company's expertise in software engineering. There, I realized the importance and the enabling factor of Computer Science. It had tangible impact far beyond than just writing code.

Service & Enabling Others. I am very thankful for the many mentoring opportunities provided by my undergrad at Virginia Tech. During my freshman year, I was mentored by a senior engineering student, and the experience was invaluable. The insights I have learned from my mentor helped me develop a holistic lifestyle balancing **education, service, and fraternity**.

I would later pay it forward by becoming a mentor myself. Passing on the tenets and values that have been taught to me, I supported and encouraged younger students to use their skills to give back to the community. I tutored and enabled student athletes to excel in their coursework, and I guided prospective college students through class selection to form an action plan for their education. For my service to the University, I was awarded **Outstanding Senior**, an honor given to one graduating ECE senior every year.

Graduate Studies. Immediately upon graduating from Virginia Tech, I started my PhD program in Computer Science and Engineering at the University of Washington (UW). I spent this summer studying data-centric systems at the warehouse-scale level working with my advisors: Luis Ceze and Mark Oskin. I surveyed emerging data-centric applications in Computer Vision and Machine Learning and decided to work on an ambitious project **to catalog the vast amounts of multimedia data in the Internet using a visual search engine**. Here, images and video are part of the search input, and media of similar visual characteristics are returned.

I am extremely grateful for the enthusiasm, mentorship, and support provided by my advisors. I am also very fortunate to receive a one-year graduate fellowship from UW. I will continue my work with visual search engines and will publish a paper by the end of the year.

(2) Relevant Background

Software-defined Radio. My first formal research experience was with the NSF Center for High-Performance Reconfigurable Computing at Virginia Tech. Dr. Athanas instilled in me that research is a **holistic** process and brilliance is not enough to succeed. He taught me that **collaboration** and **communication** would play critical parts in my career as a young scientist. Great research is done, never in isolation, but in the company of others. Leading by example, I organized weekly lab bike-rides to the Virginia countryside allowing students to collaborate outside of the confines of their cubicles.

With the support of Dr. Athanas's post-docs and graduate students, I developed a software prototype for Software-Defined Radio, a radio communication system where fixed-function hardware is replaced by reconfigurable software. This research culminated in a first-author paper and presentation at the 2013 IEEE Int. Conf. on Communications.

Teamwork and Leadership with Intel. At my internship with Intel, I learned the value of **teamwork**. My managers, Sharon and Ken, helped me reach my potential by giving me tasks that required cross-collaboration. I worked with several employees oftentimes serving as a bridge between disparate viewpoints. Even though I was just an intern, my co-workers found value in my work and treated me like a full-time employee.

I also had the opportunity to mentor and guide a new employee. I learned the value of **leadership**. Inspired by Sharon and Ken, I learned that strong leadership is cultivated through nurturing and positive relationships. On the last week of my internship, the new employee praised me for my enthusiasm and persistence throughout the on-boarding process and promised to lead by example for the next employee.

My work culminated in a methodology for validating a new power management feature for Intel's integrated graphics processing units. I was offered full-time employment after graduation.

Graphics Processing Units (GPUs). My second formal research experience was with Dr. Feng at Virginia Tech. Dr. Feng taught me the value of **tenacity**. At the time, NVIDIA released a new mechanism in their GPUs that allowed programmers to improve parallel performance at the cost of higher effort. Few had understood the implications of such a feature, and I was determined to be the first to unearth its impact. Despite many dead ends and disappointing performance figures, Dr. Feng continued to support and encourage my research. With the help of a senior graduate student, I eventually developed a novel algorithm for matrix transpose that proved the effectiveness of this new mechanism. My work culminated in the **Gold Medal Award** at the ACM Undergraduate Student Research Competition (SRC) at IEEE Supercomputing 2013.

Fast Fourier Transform (FFT). I capped off my third and final undergraduate research experience with Dr. Feng at the end of my Junior year. I developed a methodology for achieving optimal performance for the FFT even in the presence of vendor, generation, and microarchitecture variability. This work culminated in a first-author publication and presentation at the 2014 ACM Conf. on Computing Frontiers.

Evaluated as a whole, my research work was recognized by several institutions: (1) the Department of Computer Science at Virginia Tech awarded my work with **"Best Undergraduate Research"** for two consecutive years, (2) the Computing Research Association awarded my work with an **Outstanding Undergraduate Researchers Award (Honorable Mention Award)**, and finally, (3) the ACM Student Research Competition awarded my work **Third Place** at the ACM Grand Finals.

(3) Future Goals

Research Interests. Today's era of computing can no longer "throw processors at a problem." Instead, fundamental research in the co-design of the hardware and software is imperative to push performance. I will continue to work with my advisors on data-centric systems and I will contribute to the fundamental understanding of data in computing systems and architecture. The design space within the computing stack is very large, and there are many opportunities to innovate from programming models all the way down to the circuits. **If awarded, the fellowship will give me greater latitude in pursuing high-risk, high-impact ideas in Computer Architecture.**

Nurturing Collaborations. My experience with the high-performance computing and computational science community emphasized interdisciplinary collaboration. In this community, there are three main cohorts: (1) the domain scientist who is typically engrossed in the science itself, (2) the applied mathematician who develops the algorithms to solve numerical problems, and (3) the computer scientist or architect who improves the efficiency of computing systems. I will engage in **cross-cutting collaborations** that requires significant coordination between disparate scientific groups. By doing so, scientists and mathematicians can focus on their science, and I can focus on architecting computing systems to meet future application needs.

Maximizing Impact. While technical publications are important for academic success, I believe research should be gauged by its contribution in **addressing current issues**. I aim to develop software techniques that will increase operating efficiency and accelerate work flow. My yardstick for success will be determined by the number of users who find value in my line of research, the number of new applications enabled, and most importantly, the number of students, faculty, staff, and scientists whom I engage in developing these technologies.

Career Outlook. I aspire to become a professor to educate and inspire the next-generation. Because Computer Science moves at such a rapid pace, I aim to bridge the gap between theory and practice. For instance, heterogeneous parallel processing is commonplace in today's computing systems, yet current computer science curricula do not emphasize parallelism as part of the core curriculum. As a professor, I will address such examples of "**educational divide**" by working with educators to develop a curriculum that brings theory into the practice. For parallel computing, I will develop pedagogical foundations for teaching introductory computer science topics with parallelism (vs. sequential computing) at the forefront. I have already started addressing parallel computing at Virginia Tech where I led several lab-lectures using computing resources provided by NVIDIA Corporation. My experiences advocating parallel computing at the university level by student-led lectures has been met with rave reviews, and I will extend these presentations towards the K-12 audience by using the local school district in the Seattle, WA area. **If awarded, I will address the education gap between theory and practice by teaching relevant computing constructs as early as K-12**.

Intellectual Merit Criterion

Overall Assessment of Intellectual Merit

Excellent

Explanation to Applicant

Carlo worked as graphics validation engineer for Intel company for five month and also software engineer intern for Leidos. He won gold medal awards for ACM student research competition (SRC) at IEEE.ACM supercomputing conferences 2013 and third place for SRC 2014. He also received various fellowships and scholarships since 2011. As a first-year graduate student, he already published two conference papers all as the first author and one SC13 poster with best undergraduate research award. He was impressed with an ocean simulation projected onto a dome-shaped room during his internship at a naval base in New Jersey after his freshman year. His first formal research experience at the NSF Center for High-Performance Reconfigurable Computing at Virginia Tech led to a software-defined radio for changeable radio communication. He also worked as intern for Intel and designed a validating method for power management feature for Intel's integrated GPU. He also designed a new algorithm for matrix transpose and optimal FFT method under a professor at Virginia Tech. He later was admitted to the Ph.D. program of U. of Washington and decided to work on memory and networking design for scalable cataloging multimedia data using a visual search engine. Due to his education background and project experience, his research plan is well designed and convincing. The area he delves into also appears to be very promising.

Broader Impacts Criterion

Overall Assessment of Broader Impacts

Good

Explanation to Applicant

Carlo led several lab-lectures using computing resources provided by NVIDIA Corporation. He tutored student athletes to excel in their coursework, and guided prospective college students through class selection.

Summary Comments

Carlo has the education background from both ECE and CS. His various research project experiences have won him many research awards including ACM SRC awards and various fellowships. As a Ph.D. student, he would like to explore the system and architecture design for efficient multimedia data cataloging using novel memory and networking strategy. Considering the hardware aspect in order to improve the software performance is important and effective. Based on his excellent research outcome, it is very promising that he will make further significant contribution to his research area. On the other hand, his outreach activities are relatively limited.

Intellectual Merit Criterion

Overall Assessment of Intellectual Merit

Excellent

Explanation to Applicant

The applicant proposes to build a system for visual search for multimedia data (that is, serach for video clips using another clip as a search term). This is one of the most important problems that are unsolved in the areas of multimedia and information retrieval and the creation of a working retrieval system will make a high impact on the research community. To tackle the proiblem the applicant will leverage his past research experience in high performance computing (he had a paper on FFT parallelization in Supercomputing as an undergraduate student!). The components that will possibly play a significant role in the endeavor are FPGA, GPU, KD-trees, and perhaps mostly importantly feature extraction from the video. The applicant has formed an ideal group of professors who will be able to supervise his work. The faculty members unequivocally praise the applicant's high research skills, drive for research, and intelligence. His excellent academic record and the prizes he has won support the praise.

Broader Impacts Criterion

Overall Assessment of Broader Impacts

Very Good

Explanation to Applicant

If the problem proposed here is solved, it will make a great impact on the relevant research communities as well as on the society in the long run.

Summary Comments

A very fascinating research proposal from an applicant with a proven record of excellent research.

Intellectual Merit Criterion

Overall Assessment of Intellectual Merit

Excellent

Explanation to Applicant

Stellar u/g record at VaTech, including abundant research experience. Statement of how lessons were drawn from these experiences is quite compelling. And as a bonus these projects bore fruit in publications. The applicant has clearly already found a place working with three faculty at UW: two in architecture, one computer vision. His proposal for visual search is quite ambitious, and I confess some skepticism that the key advances will be at the architectural level. Nevertheless, the proposal exhibits clear and deep thought, and seems very likely to lead to signficant new technique.

Broader Impacts Criterion

Overall Assessment of Broader Impacts

Very Good

Explanation to Applicant

The project is distinctly motivated by a significant issue for computational resources on the public Internet. Success would enable many applications of broad impact. The proposal also includes sincere expressions of leadership and obligations to give back to society, as well as a few outreach suggestions.

Summary Comments

Very solid technical background and proven research productivity. This combined with a most favorable environment for making progress on an ambitious multidisciplinary project.