Ministry of Science and ICT

IITP Institute of Information & Communications Technology Planning & Evaluation

APPLICATION FORM (JOINT RESEARCH)

HIGH POTENTIAL INDIVIDUALS GLOBAL TRAINING PROGRAM)

AGREEMENT

As stated above, I submit this application form to IITP that conducts “High Potential Individuals Global Training Program” supported by Ministry of Science, ICT in South Korea. IITP may disclose the information below to the public for the purpose of providing information and matching a research partnership between your institute and a Korean university.

* IITP : Institute for Information & communications Technology Planning & Evaluation

Printed Name of Chief of Research	Kangwook Lee	Date(mm-dd-yyyy)	1/29/2020
Signature of Chief of Research	KLee		

☞ ***(Note)*** *This application is to identify the willingness to participate in this research and to find a research partnership for research institutes in Korea. Therefore, in its sole discretion, it is acceptable to contain only minimal information. (max. 3 pages)*

1. Research Title	Speeding Up Distributed AI Systems Using Deep Learning-based Codes						
2. Research Area	**A.I.**	**Big Data**	**Cloud Computing**	**Block Chain**	**AR/ VR**	**ICT/SW Convergence**	**Other ICT /SW**
	X	**X**	**X**				
3. Chief of research	Title	Assistant Professor	Contact	E-mail : kangwook.lee@wisc.edu			
	Name	Kangwook Lee		Tel : +1-608-977-1500			
4. Affiliation	Name	University of Wisconsin-Madison	Classification	(X) University ( ) Research Institute ( ) Industry ( ) ETC.			
5. Capacity for students (5 or less)	5	**Support for students *(all necessary)***	(X) Visa support (X) Research Mentoring (X) Research Space (X) Accessibility to Research equipment				

6. Research Objective

This proposal aims to usher the science of coding theory and deep learning for speeding up distributed AI systems. The distributed AI systems are characterized by a pervasive system-level noise due to a host of factors such as slow worker nodes, component failures, and communication and storage I/O bottlenecks. If left unattended, this can lead to severe system performance degradation. Recently, it has been shown that codes can speed up distributed computing systems. However, the existing solutions are based on linear codes, making them inapplicable to highly nonlinear computation tasks such as deep learning. This necessitates the need for a new nonlinear code design framework. In this project, we plan to study the combined use of coding theory and deep learning in designing nonlinear codes. As a result, we will enable the design of truly scalable distributed AI systems in the face of a diverse source of uncertainty, reclaiming the promised speed-ups.

7. Research Summary

This proposal aims to usher the science of coding theory and deep learning for speeding up distributed AI systems. Codes allow for a judiciously orchestrated redundant representation of information, offering robustness against noise. They have been used with great success to provide system reliability and fault-tolerance in many engineering applications, such as communications and storage systems, which are characterized by a variety of persistent noise.

Likewise, the distributed AI systems are characterized by a pervasive "system-level noise" due to a host of factors such as slow worker nodes, component failures, and communication and storage I/O bottlenecks. If left unattended, it leads to severe system performance degradation.

Several studies on distributed learning have consistently observed that there is a tremendous gap between the ideal and realizable distributed speedup gains when using more than tens of compute nodes. This in turn causes a drastic reduction in the pace of scientific discovery and the development, testing, and deployment of state-of-the-art machine learning models in practice.

To observe this phenomenon, we trained ResNet on CIFAR10 dataset with a varying number of distributed workers. We observed that the performance gains promised by large-scale AI systems vanish when scaling out to beyond a few tens of compute nodes.

One of the main causes of this performance degradation is the delays attributed to *straggler nodes*, i.e., nodes that are significantly slower than the average node in a distributed system. Stragglers can be attributed to the high level of heterogeneity and complexity of modern distributed systems which introduces significant delays due to resource sharing, network latency, maintenance activities, and power limits. When a distributed algorithm requires explicit synchronization between multiple tasks, as is the case for synchronous model training that is widely implemented in many modern learning frameworks, its runtime is determined by the slowest response time of the distributed tasks, which could be significantly larger than the average response time. Thus, even a single straggler node can significantly slow down the overall algorithm: this is known as the straggler problem and has been widely observed in many distributed applications.

To see whether this is indeed the cause of the performance degradation, we measured the gradient computation times of distributed workers in our distributed AI system and report the probability distribution of computation time in the figure. We can see that while most of the computation times are concentrated around the median runtime, there exist tail computation times which are up to 8x larger than the median runtime, proving the existence

	of stragglers. Recently, my group has pioneered the area of codes for distributed AI systems. In my seminal work, I propose the idea of *coded computation*, which can alleviate the straggler problem by assigning redundant jobs as per a carefully designed code. Since then, coded computation has been applied to a wide variety of settings, and there have been more than a hundred of conference/journal papers on coded computation. Recently, our work has been chosen as one of the most accessed and cited papers among those published in IEEE Transactions on Information Theory. While this new research area has rapidly grown in the past few years, all the proposed approaches are limited in their applicability. Most of the existing approaches are designed based on linear codes, making them inapplicable to highly nonlinear computation tasks such as deep learning. This precisely motivates us to raise the following fundamental research questions. Can we design nonlinear codes so that they can be used to speed up distributed AI systems involving nonlinear computation tasks? If so, how can we efficiently design such codes? As nonlinear codes are much more complicated to handcraft than linear codes, we need to develop a new code design framework. Motivated by the aforementioned research questions, our goal is **to speed up distributed AI systems with the combined use of coding theory and deep learning in designing nonlinear codes**. Our specific aims are: 1) to propose a new nonlinear code design algorithm based on a novel combined use of coding theoretic tools and deep learning algorithms, 2) to develop nonlinear codes for robust, scalable distributed AI systems with nonlinear operations, 3) to identify concrete cases/applications where the proposed solutions can achieve 10x—100x speed-ups and/or resource savings, compared to the existing solutions, 4) to implement the proposed solution as a plugin for existing distributed AI frameworks, and 5) to evaluate the practical performance of our solution with real-world system logs collected from large-scale distributed AI systems. This research project will be highly collaborative: The main collaborators will be Prof. Dimitris Papailiopoulos (University of Wisconsin-Madison, ECE) and Prof. Shivaram Venkataraman (University of Wisconsin-Madison, CS). The visiting students will have a chance to interact with all of the collaborators.
8. Need for funding from Korean government	100M KRW + student funding
9. Request for Korean Universities	- The selection of students studying abroad should be conducted after mutual consultation, and please cooperate as much as possible to prepare for VISA. - Strong backgrounds in **mathematics and deep learning** are required.