

이력서

작성일 : 2023. 09. 10

성명	윤대건	영문성명	Daegun Yoon
생년월일	1994.02.04	E-mail	slashxp@naver.com

학력사항	구분	입학년월	졸업년월	학교명	전공	졸업구분
	고등학교	2009.03	2012.02	영동일고등학교	자연계	졸업
	대학교	2013.03	2018.08	아주대학교	소프트웨어및컴퓨터공학전공	졸업
	대학원 (석박사통합)	2018.09	2024.02	아주대학교	인공지능전공	졸업예정
		졸업논문	ExDyna: Preserving Near-Optimality of Gradient Sparsification for Data-Parallel Distributed DNN Training			
		LAB.	병렬 및 분산 처리 연구실, 지도교수: 오상윤, 홈페이지: https://wise.ajou.ac.kr/			

주요 수행 프로젝트

회사 / 기관명	기간	프로젝트명
한국연구재단	2022.03 ~ 현재	BK21+ 4단계, 대규모 분산 학습 가속을 위한 통신 최적화 기법의 계산/통신 비효율성 개선
[담당업무 및 주요성과] 분산 딥러닝 Gradient sparsification 계산/통신 비용 최적화 기법 연구 <ul style="list-style-type: none">- NeurIPS, MLSys 등에서 발표된 State-of-the-art Gradient sparsification 기법들의 한계점 분석- 대규모 분산 학습을 위해 높은 확장성을 목표로 하는 고성능 Gradient sparsification 기법 설계- SOTA, 제안 기법 구현 및 성능 검증 (PyTorch 기반)- SCI급 학술지 논문 주저자 1편 채택- BK21+ CS 분야 우수국제학술대회 논문 주저자 제출 (1편 채택, 1편 심사 중 - Rebuttal 완료, 1편 2023.11 초 제출 예정)		

회사 / 기관명	기간	프로젝트명
KISTI (한국과학기술정보연구원)	2022.03 ~ 2022.10	거대 딥러닝 모델 분산 학습 속도 향상을 위한 샤드 데이터 병렬화 메모리 활용 최적화 및 통신 스케줄링 기법
[담당업무 및 주요성과] 고성능 GPU 클러스터 기반의 슈퍼컴퓨터에서 DeepSpeed와 FairScale 벤치마크 및 새		

로운 Sharded data parallel 분산 학습 최적화 기법 설계

- Multidimensional benchmark 수행
 - 프레임워크={DeepSpeed: ZeRO 0, 1, 2, 3}, {FairScale: ZeRO 0, 2, 3}
 - 모델={이미지 생성 모델 DALL-E: original, 1B}, {자연어 처리 모델 T5: small, base, large}
 - GPU={V100 32GB NVLink: 1, 2, 4, 8대}, {A100 80GB NVLink: 1, 2, 4, 8대}
- 벤치마크 결과, 논문, 소스코드 분석을 통해 ZeRO의 메모리 관리 문제점 도출
- ZeRO의 메모리 비효율성을 개선하기 위한 파라미터 공유 스케줄 최적화 기법 설계

회사 / 기관명	기간	프로젝트명
KISTI (한국과학기술 정보연구원)	2019.03 ~ 2019.10	초거대규모 CPU 기반 환경에서 딥러닝 응용의 12페타플롭스 병렬확장성 연구
[담당업무 및 주요성과] 대규모 CPU 클러스터에서 딥러닝 플랫폼 운용을 위한 최적화 연구 <ul style="list-style-type: none"> - 분산 딥러닝의 학습 속도 및 정확도 향상을 위한 레이어 별 통신, 비동기 통신 기반 SGD 등 분산 학습 최적화 기법 연구 		

회사 / 기관명	기간	프로젝트명
KISTI (한국과학기술 정보연구원)	2018.09 ~ 2018.10	거대규모 인공지능경망 학습 및 검증, 응용연구 활용을 위한 슈퍼컴퓨팅 환경에서의 병렬 플랫폼 최적화 연구
[담당업무 및 주요성과] 대규모 CPU 클러스터에서 딥러닝 응용 실행 시 제한요소 및 성능 저하 요소 도출 연구 <ul style="list-style-type: none"> - 대규모 데이터셋 벤치마크 결과를 분석하여 CPU 기반 딥러닝 응용에 적합한 메모리 및 네트워크 최적화 요소 도출 		

회사 / 기관명	기간	프로젝트명
삼성디스플레이	2023.01 ~ 현재	고효율 HPC Job Scheduling 알고리즘 개발
[담당업무 및 주요성과] 삼성디스플레이의 해석 시뮬레이션이 구동되는 데이터 센터의 고성능컴퓨팅 환경에서 효율적인 작업 스케줄링을 가능하게 하는 알고리즘 개발 <ul style="list-style-type: none"> - 실제 데이터 센터의 컴퓨팅 리소스를 물리 코어 단위의 클러스터의 토폴로지로 표현하여 현업에서의 작업 스케줄링에 문제가 되는 요소 분석 - 데이터 센터에서 구동된 시뮬레이션들의 작업 로그를 데이터로 머신 러닝 및 데이터 마이닝을 수행하여 작업 스케줄링에 큰 영향력을 행사하는 요소 도출 - 실제 데이터 센터의 클러스터 토폴로지와 작업 로그를 기반으로 하여 실제 운용 상황에서의 리소스 점유 		

<ul style="list-style-type: none"> 을 및 작업 대기시간을 재현하는 스케줄링 시뮬레이터 개발 (C++ 기반) - 개발한 스케줄링 시뮬레이터에서 다양한 스케줄링 정책 (FCFS, SJF 등)을 적용하여 확인한 결과와 각 정책의 장단점을 분석하여 삼성디스플레이 데이터 센터에 특화된 스케줄링 알고리즘 설계 및 개발
--

회사 / 기관명	기간	프로젝트명
한국연구재단	2020.09 ~ 2022.02	BK21+ 4단계, NVIDIA GPU 아키텍처의 Warp 기반 병렬성을 고려한 고성능 그래프 알고리즘
[담당업무 및 주요성과] GPU 기반 고성능 BFS 알고리즘 개발 <ul style="list-style-type: none"> - ASPLOS, SC, HPDC, ATC에서 발표된 그래프 프레임워크 및 BFS 알고리즘들의 한계점 분석 - 그래프 알고리즘 고성능화 및 병렬화를 위해 활용해야 하는 GPU 아키텍처 요소 도출 - Warp 기반 고성능 Direction-optimizing BFS 알고리즘 설계 - 기존, 제안 기법 구현 및 성능 검증 (C++, CUDA 기반) - SCI급 학술지 논문 주저자 2편 채택 		

전공 소개서

<p>[연구 개요]</p> <p>저는 아주대학교 일반대학원 인공지능학과 석박사통합과정으로 2018년 9월부터 현재까지 병렬 및 분산 처리 분야의 다양한 주제로 연구를 수행해왔습니다. 학위 과정 중 1) 분산 시스템의 부하 균형을 위한 리소스 사용량 관리 및 핫 스팟 완화, 2) NVIDIA GPU 아키텍처의 Warp 기반 병렬성을 고려한 고성능 그래프 알고리즘, 3) 대규모 분산 학습 가속을 위한 통신 최적화 기법의 계산/통신 비효율성 개선의 순서로 연구를 수행했으며 연구 3은 제 대표 연구 주제입니다. 연구 1과 2를 수행하면서 얻은 경험과 인사이트를 기반으로 연구 3에서의 계산 및 통신 비용을 줄이는 확장성 있는 분산 학습 통신 최적화 기법을 설계할 수 있었고, 해당 기법은 부하 분산과 GPU 코어 병렬성을 활용하여 큰 성능 향상을 달성했습니다. 이러한 결과로부터 AI 기술을 활용하는 다양한 과학기술 응용 개발에 필요한 대규모 분산 학습의 가속화 및 비용 절감으로 산업에서의 가치를 창출하는 데에 기여할 수 있을 것으로 기대하고 있습니다.</p> <p>연구 1) 분산 시스템의 부하 균형을 위한 리소스 사용량 관리 및 핫 스팟 완화 (2018.09 ~ 2020.02)</p> <p>[연구 배경]</p> <ul style="list-style-type: none"> - 콘텐츠 기반 출판/구독 형식의 분산 메시징 시스템에서는 사용자가 전달받기를 원하는 정보가 구독의 형태로 서버에 저장되며, 해당하는 정보가 출판되면 서버에서 사용자에게 전송 - 특정 정보에 대한 수요가 많아지면 해당 서버들의 부하가 증가하여 정보에 대한 수요의 불균형으로 핫 스팟 발생 - 수요가 집중되는 정보의 구독 복제본을 여러 서버로 분산하여 핫 스팟을 완화할 수 있지만, 복제본의 수가 많아질수록 서버들의 메모리 사용량과 복제본을 담당하는 스레드의 수가 증가하여 비효율적인 리소스 점유와 성능 감소 초래
--

[연구 내용 및 결과]

- 본 연구에서는 분산 시스템 내의 서버들의 리소스 점유 상태와 각 정보에 대한 수요를 실시간으로 모니터링하여 서버들에 배포되는 복제본의 수를 동적으로 조절함으로써 메모리/CPU 사용량 관리와 핫스팟 완화 간의 균형을 최적화하는 기법 제안
- 본 연구를 수행한 결과로 정보의 복제본을 감소시켜 리소스 오버헤드가 감소하고 부하 분산으로 핫스팟을 완화하는 효과를 달성
- 본 연구를 통해 얻은 경험을 토대로 분산 환경의 HPC 클러스터에서 실행되는 병렬/분산 응용들의 리소스 사용 효율성 개선과 성능 향상을 위한 기술 개발에 기여할 수 있을 것으로 기대

[관련 SCI급 주저자 논문 실적]

- [1] Daegun Yoon, Zhetao Li, Sangyoon Oh, "Balanced content space partitioning for pub/sub: a study on impact of varying partitioning granularity", The Journal of Supercomputing, Apr. 2021. (IF=3.3)
- [2] Daegun Yoon, Gyudong Park, Sangyoon Oh, "Exploring a system architecture of content-based publish/subscribe system for efficient on-the-fly data dissemination", Concurrency and Computation: Practice and Experience, Nov. 2020. (IF=2.0)

연구 2) NVIDIA GPU 아키텍처의 Warp 기반 병렬성을 고려한 고성능 그래프 알고리즘 (2020.03 ~ 2022.02)

[연구 배경]

- 수많은 그래프 알고리즘들의 기본형인 BFS는 탐색 방법에 따라 Push 또는 Pull 기반의 알고리즘으로 구분되며, 각 탐색 방법은 그래프의 형태와 워크로드 상태에 따라 성능이 상이
- Direction-optimizing BFS는 Push와 Pull 중 현재 상태에 더 적합하다고 판단되는 방법을 선택하여 그래프를 탐사하도록 하는 방법
- 기존 기법들은 실제 워크로드를 고려하지 않고 Push/Pull 선택을 그래프 데이터마다 튜닝한 조건에 의지하여 다양한 그래프 데이터에 대해 일관성 있는 고성능을 달성하기 어려움
- 기존 기법들은 Push/Pull의 설계 및 구현에서 GPU 기반 알고리즘에서 Atomic operation이 코어들의 병렬성으로 얻는 성능 향상을 저하시킨다는 점과 Memory/Branch divergence 발생 가능성을 충분히 고려하지 않아 성능 저하 발생

[연구 내용 및 결과]

- 본 연구에서는 각 탐색 방법을 GPU 아키텍처에 적합하게 설계하여 병렬성을 최대 활용할 수 있도록 하고, 탐색 방법을 워크로드 상태를 고려하여 선택하도록 하는 고도화된 알고리즘 제안
- 최근 Iteration에서의 Frontier (이전 Iteration에서 방문된 Vertex) 개수 변화를 모니터링하고 Frontier의 Variation과 Convexity에 기반하여 워크로드의 규모와 상태 판단
- 제안 기법은 워크로드의 규모가 클 때는 Atomic-free한 탐색 알고리즘을 선택하고, 규모가 작을 때는 Atomic operation 기반의 알고리즘을 선택
- Atomic-free한 알고리즘은 Warp 기반의 병렬성을 극대화하기 위해 Memory/Branch divergence 발생을 최소화하도록 설계
- 본 연구를 수행한 결과로 다양한 그래프 데이터들에서 기존 기법들에 비해 그래프 처리 성능이 크게 향상되었으며 특정 그래프 데이터에 종속되는 튜닝 없이 일관성 있는 고성능 달성
- 본 연구를 통해 얻은 GPU 기반의 병렬 프로그래밍 경험으로 고성능 GPU 클러스터에서의 거대모델 학습 최적화에 기여할 수 있을 것으로 기대

[관련 SCI급 주저자 논문 실적]

- [3] Daegun Yoon, Minjoong Jeong, Sangyoon Oh, "WAVE: designing a heuristics-based three-way breadth-first search on GPUs", The Journal of Supercomputing, Nov. 2022. (IF=3.3)
- [4] Daegun Yoon, Sangyoon Oh, "SURF: Direction-Optimizing Breadth-First Search Using Workload State on GPUs", Sensors, Jun. 2022. (IF=3.9)

연구 3) 대규모 분산 학습 가속을 위한 통신 최적화 기법의 계산/통신 비효율성 개선 (2022.03 ~ 현재)

[연구 배경]

- Gradient sparsification은 정보를 손실시키는 Lossy algorithm의 일종으로, Data-parallel 분산 학습에서 각 워커가 자신의 Mini-batch에서의 역전파로 계산된 로컬 Gradient tensor의 모든 Gradient를 All-reduce 하지 않고 극히 일부 Gradient만 통신으로 교환 (Sparse All-reduce)하는 통신 최적화 기법
- NeurIPS, MLSys 등에서 발표된 State-of-the-art Gradient sparsification 기법들의 논문들에서는 Sparse All-reduce가 적용된 SGD의 수렴을 이론적으로 설명하는 데에 많은 노력을 기울였지만, 여기에는 많은 이론적 가정이 포함되어 있고 알고리즘 측면에서의 한계로 실제 분산 학습에 실용적으로 적용되기 어려움
- 본 연구는 기존 기법들에서 발생하는 비용들을 크게 줄여 실제 산업의 제품들에 탑재되는 딥러닝 모델 개발 비용을 줄이는 데에 활용될 수 있도록 실용성에 초점을 둠
- 본 연구에서는 Gradient sparsification의 정보 손실로 인한 모델의 정확도 손실을 최소화하면서 최적화를 위한 추가적인 오버헤드 없이 계산 및 통신 비용을 크게 줄이는 확장성 있는 분산 학습을 가능하게 하는 새로운 기법을 제안하는 것을 목표로 함

[연구 내용 및 결과]

- 본 연구에서는 SOTA Gradient sparsification 기법들의 한계점을 알고리즘 측면에서 상세히 분석하여 계산 및 통신 상 제약 사항으로 분류하고, 이를 해결하기 위해 새로운 형태의 Gradient sparsification 기법인 DEFT를 제안
- DEFT는 기존 기법들에서 발생하는 계산 및 통신 제약 사항으로 분류된 총 6가지 한계점 (표 1)을 극복하는 것을 목표로 설계

표 1 기존 Gradient sparsification 기법들과 제안 기법 DEFT의 계산 및 통신 측면 장단점

Sparsifier	Communication-constraints			Computation-constraints		
	Gradient build-up	Unpredictable density	Hyperparameter tuning	Worker idling	Gradient selection cost	Additional overhead
Top-k	Yes	Yes	No	No	Very high	No
CLT-k [R1]	No	No	No	Yes	Very high	No
Hard-threshold [R2]	Yes	Yes	Yes	No	Very low	No
SIDCo [R3]	Yes	Yes	No	No	Very low	Very high
DEFT (Proposed)	No	No	No	No	Low	Very low

- DEFT는 Sparse All-reduce의 대상이 될 Gradient를 선택하는 과정을 병렬화하고 학습에 참여하는 워커들 간의 부하 균형을 맞추는 새로운 Gradient sparsification 기법이며, 그림 1은 4단계의 과정으로 구성된 DEFT의 개요를 나타냄

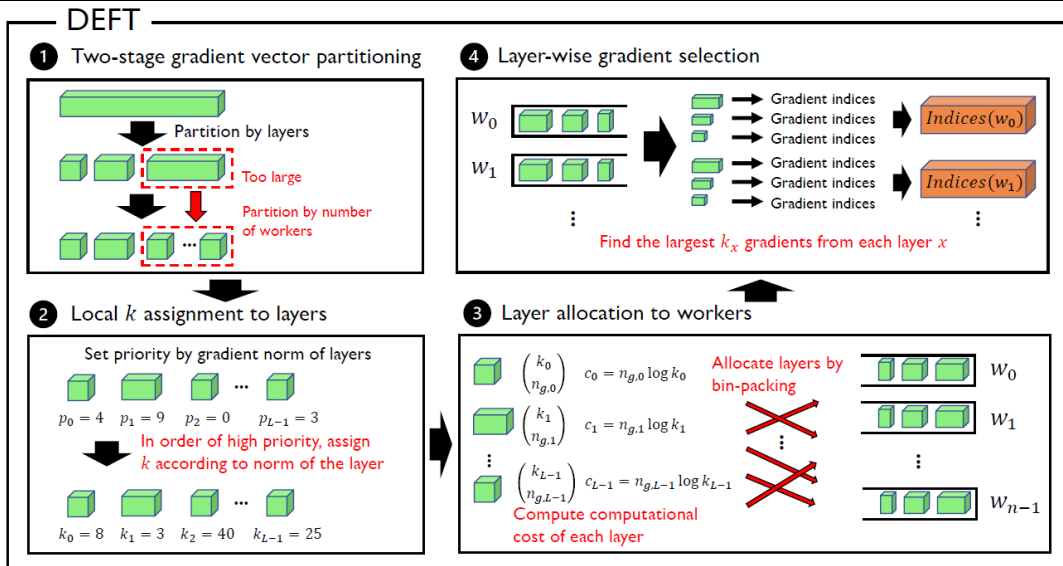
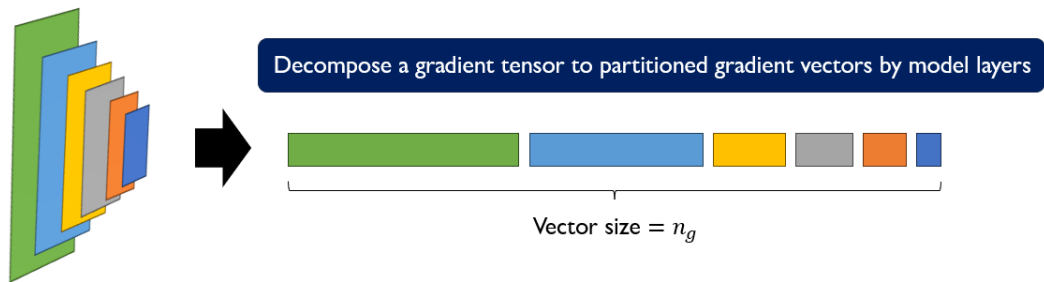


그림 1 제안 기법 DEFT 개요

- 과정 1: Two-stage gradient vector partitioning (그림 2)
 - DNN 모델의 Gradient tensor를 1차원 벡터로 변환한 후 레이어 별로 분할
 - 분할된 벡터들 중 Threshold (= 모델의 Gradient 개수 / 워커의 수)보다 크기가 큰 벡터들은 워커의 수만큼 다시 분할

[Stage 1: Coarse-grained partitioning by layers]



[Stage 2: Fine-grained partitioning by threshold] = $\frac{n_g}{n}$

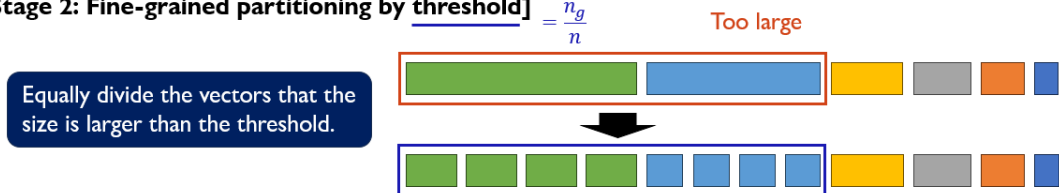
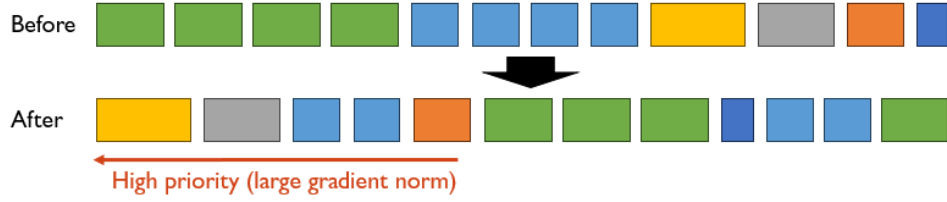


그림 2 Two-stage gradient vector partitioning

- 과정 2: Local k assignment to layers (그림 3)
 - DNN 모델의 레이어 (분할된 벡터)들을 레이어의 Gradient norm 기준으로 정렬하여 값이 큰 레이어에 높은 우선순위를 부여
 - 높은 우선순위의 레이어들에서 더 많은 Gradient를 선택할 수 있도록 더 큰 k (선택되는 Gradient 개수)를 할당

2a Prioritize layers by their gradient norms



2b Assign k to each layer in priority order

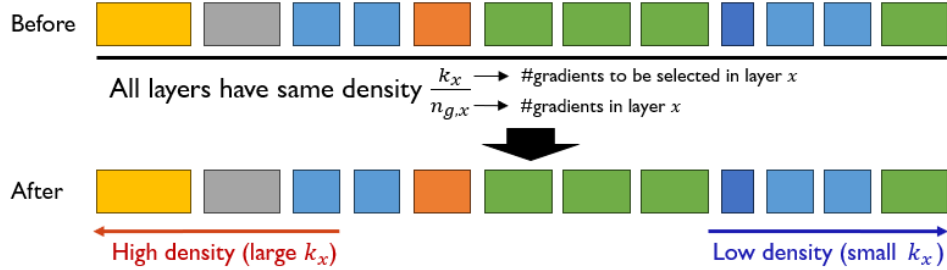
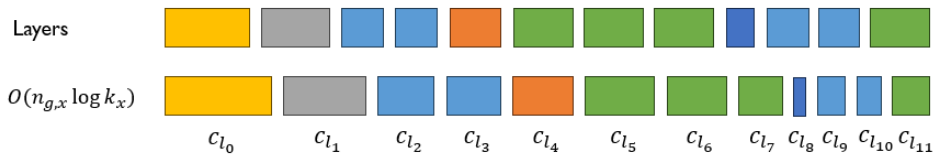


그림 3 Local k assignment to layers

- 과정 3: Layer allocation to workers (그림 4)
 - Gradient selection을 위한 Top-k operation의 계산 복잡도 $O(n \log k)$ 를 기준으로 각 레이어의 계산 비용을 계산
 - 워커들 간 Gradient selection 부하 균형을 고려하기 위해 각 레이어의 계산 비용을 기준으로 Bin-packing 알고리즘을 이용하여 레이어들을 워커들에게 할당

3a Compute the computational cost of each layer by formula $O(n_{g,x} \log k_x)$



3b Allocate layers to workers by c_x using bin-packing algorithm

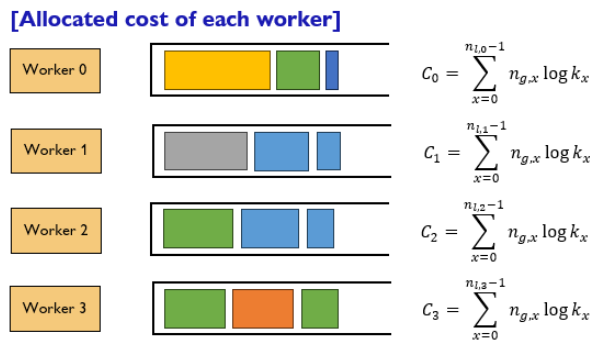
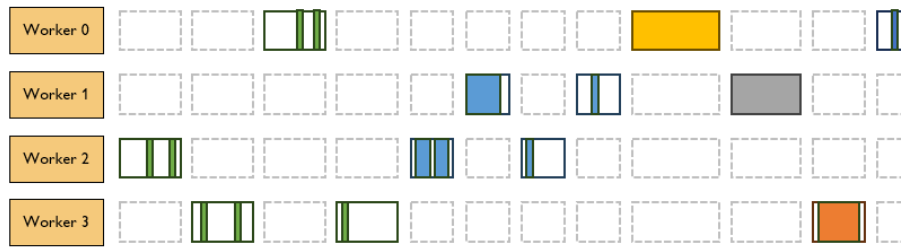


그림 4 Layer allocation to workers

- 과정 4: Layer-wise gradient selection (그림 5)
 - 각 워커는 자신이 할당받은 레이어들에서 Gradient selection을 수행
 - 병렬화된 Gradient selection으로 인해 계산 비용 크게 감소
 - 통신 트래픽을 증가시키는 문제인 Gradient build-up을 방지하여 Sparse All-reduce 통신 비용 크게 감소

[Selected gradients by workers]



[Sparsified gradients]

All-gather

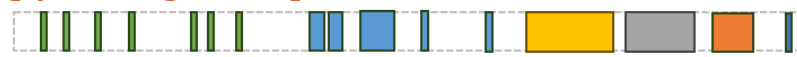
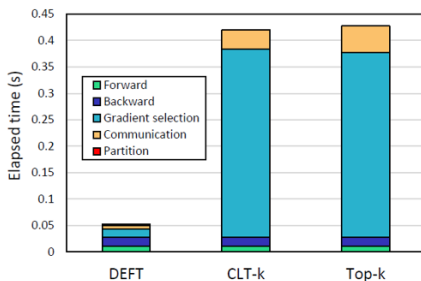
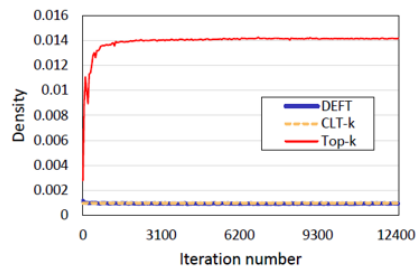


그림 5 Layer-wise gradient selection

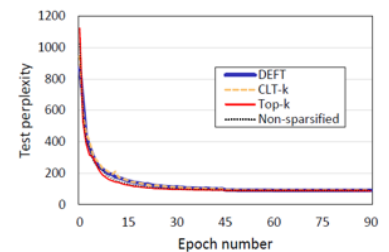
- 실험 환경
 - 하드웨어: 최대 8 노드, 각 노드는 4대의 NVIDIA Tesla V100 GPU 탑재 (CUDA 10.1)
 - 프레임워크: PyTorch 1.5, NCCL 2.4, Open MPI 4.0.5
 - 응용: Language modeling (LSTM & WikiText-2) (논문에서는 더 다양한 응용으로 실험)
 - 비교군: DEFT, Top-k, CLT-k, Non-sparsified
- 주요 실험 결과
 - Overall performance (그림 6a): 한 번의 Training iteration에서 소요된 Wall-clock-time을 측정한 것으로, DEFT는 계산 및 통신 비용이 크게 감소하여 기존 기법들에 비해 크게 향상된 학습 속도를 보임
 - Sparsification performance (그림 6b): DEFT는 Gradient build-up 등 통신 트래픽을 증가시키는 요소들을 방지하여 사용자가 요구한 수준의 통신 트래픽을 유지
 - Convergence performance (그림 6c): DEFT는 모델 업데이트에 유의미한 영향을 줄 수 있는 레이어 (Gradient norm이 큰 레이어)에서 더 많은 Gradient를 선택하여 기존 기법들과 비슷한 수준 수렴 성능을 보임
 - Scalability: 그림 7a는 DEFT의 병렬화된 Layer-wise gradient selection이 기존의 병렬화되지 않은 Top-k gradient selection에 비해 달성하는 속도 향상 (Speedup)이 워커의 수가 늘어날수록 더 증가하는 결과를 나타내며, 그림 7b는 워커의 수가 늘어나도 모델의 최종 정확도가 비슷한 수준으로 유지되는 결과를 나타냄



(a) Overall performance

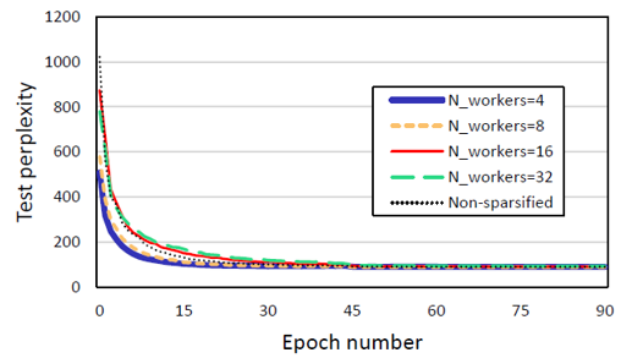
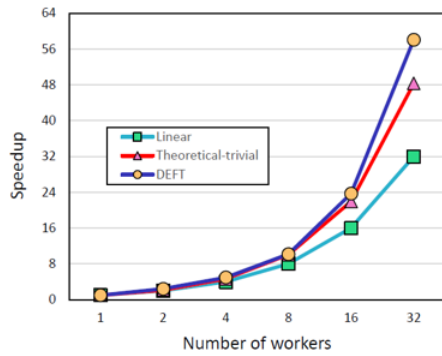


(b) Sparsification performance



(c) Convergence performance

그림 6 DEFT와 기존 기법들의 성능 평가



(a) 계산 비용 측면에서의 DEFT의 확장성

(b) 정확도 유지 측면에서의 DEFT의 확장성

그림 7 DEFT의 확장성 평가

- 기대 효과
 - 본 연구를 수행한 결과로 Sparse All-reduce에 필요한 계산/통신 비용을 크게 감소시켜 Gradient sparsification을 분산 학습에 실용적으로 적용할 수 있을 것으로 기대
 - 모델의 정확도 손실을 최소화하면서 비용 효율적인 Gradient sparsification을 거대 모델의 고성능 GPU 클러스터 기반 대규모 분산 학습에 적용하여 학습을 가속화함으로써 AI 기술을 활용하는 고성능 과학 기술 응용의 개발 비용 감소 및 품질 개선에 기여할 수 있을 것으로 기대

[관련 SCI급 주저자 논문 실적]

- [5] Daegun Yoon, Sangyoon Oh, "DEFT: Exploiting Gradient Norm Difference between Model Layers for Scalable Gradient Sparsification", 52nd International Conference on Parallel Processing (ICPP 2023), Aug. 2023. (BK21+ CS 분야 우수국제학술대회 인정 Impact Factor 2)
- [6] Daegun Yoon, Minjoong Jeong, Sangyoon Oh, "SAGE: toward on-the-fly gradient compression ratio scaling", The Journal of Supercomputing, Feb. 2023. (IF=3.3)

[참고문헌]

- [R1] Chia-Yu Chen, Jiamin Ni, Songtao Li, Xiaodong Cui, Pin-Yu Chen, Xiao Sun, Naigang Wang, Swagath Venkataramani, Vijayalakshmi Viji Srinivasan, Wei Zhang, Kailash Gopalakrishnan, "Scalecom: Scalable sparsified gradient compression for communication-efficient distributed training", 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Dec. 2020.
- [R2] Atal Sahu, Aritra Dutta, Ahmed M Abdelmoniem, Trambak Banerjee, Marco Canini, Panos Kalnis, "Rethinking gradient sparsification as total error minimization", 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Dec. 2021.
- [R3] Ahmed M Abdelmoniem, Ahmed Elzanaty, Mohamed-Slim Alouini, Marco Canini, "An efficient statistical-based gradient compression technique for distributed training systems", 4th Conference on Machine Learning and Systems (MLSys 2021), Apr. 2021.

연구실적

요약	SCI급 주저자 논문 : 6편
----	------------------

	SCI급 공저자 논문 : 2편 기타 논문 : 15회 (국내: 10회, 해외: 5회) 특허 : 2건 (국내: 2건, 해외: 0건)
--	--

[SCI급 주저자 논문] 6편

논문명	학회/저널명	Impact Factor
DEFT: Exploiting Gradient Norm Difference between Model Layers for Scalable Gradient Sparsification	52nd International Conference on Parallel Processing (ICPP 2023) [BK21+ CS 분야 우수국제 학술대회 인정 Impact Factor 2]	2.0
SAGE: toward on-the-fly gradient compression ratio scaling	The Journal of Supercomputing	3.3
WAVE: designing a heuristics-based three-way breadth-first search on GPUs	The Journal of Supercomputing	3.3
SURF: Direction-Optimizing Breadth-First Search Using Workload State on GPUs	Sensors	3.9
Balanced content space partitioning for pub/sub: a study on impact of varying partitioning granularity	The Journal of Supercomputing	3.3
Exploring a system architecture of content-based publish/subscribe system for efficient on-the-fly data dissemination	Concurrency and Computation: Practice and Experience	2.0

[SCI급 공저자 논문] 2편

논문명	학회/저널명	Impact Factor
AMBLE: Adjusting Mini-Batch and Local Epoch for Federated Learning with Heterogeneous Devices	Journal of Parallel and Distributed Computing	3.8
Mitigating Cold Start Problem in Serverless Computing with Function Fusion	Sensors	3.9

[기타 논문] 15편 (국내: 10편, 해외: 5편)

논문명	저자구분	학회/저널명	국내/해외
연합학습 기법들의 성능평가를 지원하는 이기종 기반의 실험 플랫폼 설계	공저자	2023 한국통신학회 하계학술대회	국내
Can hierarchical client clustering mitigate the data heterogeneity effect in federated learning?	공저자	5th IPDPS Workshop on Scalable Deep Learning over Parallel and Distributed Infrastructure (ScaDL	해외

		2023)	
연합학습에서의 보안 취약점 분석	공저자	2023 한국통신학회 동계학술대회	국내
Empirical Analysis on Top-k Gradient Sparsification for Distributed Deep Learning in a Supercomputing Environment	주저자	8th International Conference on Next Generation Computing (ICNGC 2022)	해외
전송망의 라우팅 성능 개선을 위한 성능 지표 분석 기반 정책 엔진 설계	주저자	한국통신학회 논문지	국내
SDN 정책엔진의 사용자 모듈을 위한 분석 요청 정의 언어	공저자	한국통신학회 논문지	국내
재난 대응 기계학습 모델의 Data Drift 문제에 대한 MLOps 기반 대응 기법	공저자	2022 한국차세대컴퓨팅학회 춘계학술대회	국내
Traversing Large Road Networks on GPUs with Breadth-First Search	주저자	7th International Conference on Next Generation Computing (ICNGC 2021)	해외
전송망 성능 개량을 위한 정책 엔진 인터페이스 설계	주저자	2021 한국군사과학기술학회 종합학술대회	국내
Imitation learning for VM placement problem using demonstration data generated by heuristics	공저자	17th International Conference on Data Science (ICDATA 2021)	해외
Software-Defined Network에서의 Conflict Resolution을 위한 정책엔진 구조 및 전략 분석	주저자	2020 한국통신학회 하계종합학술발표회	국내
CPartition: a Correlation-Based Space Partitioning for Content-Based Publish/Subscribe Systems with Skewed Workload	주저자	7th International Conference on Big Data and Smart Computing (BigComp 2020)	해외
동종 운영체제 환경에서의 가상 머신 마이그레이션 성능 분석	주저자	2018 한국정보과학회 한국컴퓨터종합학술대회	국내
온라인 뉴스, 댓글, 사용자 특성 파악을 위한 감정 분석 시스템 제작	주저자	2018 한국통신학회 동계종합학술발표회	국내
가상 AUTOSAR Platform 상에서의 Traction Control System 설계 및 시뮬레이션 방법	공저자	2016 한국정보과학회 한국컴퓨터종합학술대회	국내

[특허] 2건 (국내: 2건, 해외: 0건)

특허명	발명자	등록일 (출원일)	등록번호 (출원번호)	국내/해외 구분
워크로드 규모 분석에 따른 적응형 그 래프 탐색 장치 및 방법	오상윤, 윤대 건	2023.06.07 (2023.01.12)	1025423110000 (1020230004444)	국내
이벤트 공간 분할 방법 및 장치, 컴퓨 터 판독 가능한 기록 매체 및 컴퓨터 프로그램	박민호, 오상 윤, 윤대건, 함재현	2022.07.29 (2021.06.10)	1024287490000 (1020210075518)	국내