<주요 Q&A>

초거대 AI 연구를 위한 기반 기술 이해

Q1. 초거대 AI 연구로 나누는 기준은 데이터 양 등 어떤 기준으로 나뉘게 되는지요? 수집된 데이터를 조합하고 연관성을 추출하기 위한 방안은 어떻게 되는지요?

특별한 기준이 있다기보다는 Billion 규모의 파라미터를 가져, 단일 노드로는 학습이 불가능한 모델들을 초거대 AI 연구로 말하고 있습니다. 현재는 자연어에서 이러한 연구가 활발히 진행 중이며, 해당 논문/연구에서 제안된 방법을 기반으로 데이터를 활용하고 있습니다.

Q2. model parallelism을 효율적으로 활용하는 방안에 대해서 문의드립니다

HW와 SW 모두 중요하며, 우선 HW는 노드 내에서의 효율적인 GPU통신을 위한 NVLink/NVSwitch, 노드 간의 효과적인 통신을 위한 Infiniband 구성이 중요한 요소입니다. 또한 SW적으로도 효과적인 collective 연산을 위한 NCCL 등을 활용하여 효과적인 communication을 수행하게 됩니다.

Q3. 데이터 통신 성능을 향상시키려는 경우 distributed computing을 활용하는 방안에 대해서 질문드립니다

발표에서 언급되고 있는 SW/HW 요소가 모두 분산학습을 위한 핵심적인 요소입니다.

Q4. GPUDirect 기능이 활성화되면서 latency 와 thruougput은 좋아지지만 보안은 약해지는 것이 아닌지 궁금합니다.

GPUDirect에서도 특정 Process에서 접근할 수 있는 Memory Address는 제약이 있으며, 또한 NVIDIA에서 GPU 클러스터의 보안은 성능과 더불어 가장 중요하게 고려하고 있는 요소입니다. 최근에는 DPU를 활용한 Morpheus 등을 통해서 Security측면을 강화하려는 시도도 꾸준히 진행되고 있습니다.

Q5. GPU통신을 위한 기계적 개선 포인트는 없나요?

기존 PCIe를 통한 한계를 극복하기 위해 NVLink/NVSwitch를 NVIDIA에서 개발했고 현재 사실상의 표준으로 활용되고 있습니다. 또한, 노드 간 연결 역시 Ethernet 대비 훨씬 효과적인 Latency, RDMA 등의 장점을 갖고 있는 Infiniband 역시 중요한 요소입니다.

Q6. 연구실 GPU 클러스터를 구성하려고 하는데 조언 부탁드립니다.

현재 AI/DL 연구에서 가장 효과적인 GPU는 A100입니다. A100 4개로 구성된 Station형태의 DGX-A100 Station, 그리고 A100 8장으로 구성된 DGX-A100등을 reference architecture로 활용해주시면 좋을 것 같으며, Cluster-level에서는 DGX Pod Whitepaper등을 참조해주시면 됩니다.

Q7. GPT-3를 Nvidia GPU로 추론할 때, 충분한 GPU를 쓸 경우 서비스 턴어라운드 시간은 몇 ms내에 가능한가요? 분산작업에 필요한 최소한의 Latency가 얼마인지 궁금한 것입니다.

최근 GTC 키노트에서 GPT3보다 3배정도 큰 530B모델에 대해서 2대의 DGX Node를 통해 0.5초만에 Inference하는 benchmark가 공개된 바 있습니다.

Q8. tensor 퍼러럴 이 가장 효과적으로 작동하기 위해서는 gpu가 몇개가 필요한지 궁금합니다.

각각의 Configuration들은 Model의 크기에 따라 달라지게 되며, 대표적인 모델들에 대해서는 https://github.com/NVIDIA/Megatron-LM 페이지에서 효과적인 configuration값들을 제공하고 있습니다.

Q9. 일반 사용자의 gpu로도 페럴리즘 구현이 되나요?

가능합니다. 다만 기존 PCIe 버스, Ethernet을 기반으로한 Communication은 아무래도 NVLink/Infiniband를 활용한 communication에 비해서 성능적으로는 한계가 있을 수 있습니다.

Q10. GPU의 연결성이 증가하면서 기계학습 쪽에서 CPU의 중요성이 감소한다고 생각합니다. 혹시 Cuda를 쓴 학습에서는 CPU의 병목이 거의 없어진다고 보면 될까요?

기본적으로 해당 병목들을 줄이는 것이 GPU최적화의 중요한 포인트이며, NVIDIA는 CPU-GPU-Network-Storage 등 모든 병목현상을 최소화하기 위한 Full-stack 관점에서의 최적화를 진행하고 있습니다. 이번 세션에서 언급된 GPUDirect 기술 혹은, CPU-GPU간의 Bottleneck을 최소화하기 위한 Arm기반 CPU인 Grace 등이 대표적인 예시가 되겠습니다.

Q11. 초거대 인공지능 연구를 하는 데 있어서 NLP을 효율적으로 활용할 수 있는 방안에 대해서 질문드립니다

최근 들어 국내에도 초거대 NLP를 기반 모델의 활용에 대한 다양한 session들이 발표된 적이 있습니다. 해당 세션들을 참조해주시면 감사하겠습니다.