**Deep-Speed:**

* Ref:

1. <https://velog.io/@seoyeon96/DeepSpeed%EB%A1%9C-%ED%81%B0-%EB%AA%A8%EB%8D%B8-%ED%8A%9C%EB%8B%9D%ED%95%98%EA%B8%B0>
2. <https://velog.io/@seoyeon96/%EB%A6%AC%EC%84%9C%EC%B9%98-%ED%9A%A8%EC%9C%A8%EC%A0%81%EC%9D%B8-%EB%B6%84%EC%82%B0-%ED%95%99%EC%8A%B5%EC%9D%84-%EC%9C%84%ED%95%9C-DeepSpeed-ZeRO>
3. https://medium.com/daangn/pytorch-multi-gpu-%ED%95%99%EC%8A%B5-%EC%A0%9C%EB%8C%80%EB%A1%9C-%ED%95%98%EA%B8%B0-27270617936b

* 모델의 training & inference 속도를 빠르고 효율적으로 할 수 있도록 도와주는 Microsoft에서 만든 딥러닝 최적화 라이브러리 🡪 여러 GPU를 병렬적으로 사용할 때, 효율성 증가
* 이전 학습 방식
  + CPU
  + Single GPU
  + 1 Node, Multi GPU 🡪 **분산 학습을** 해야함
    - 분산학습에서는 데이터는 분산되지만 모델을 학습시키기 위한 값(Optimizer State, Gradient, Parameter)는 각 GPU마다 복제해서 가지고 있어야했다. 복제된 파라미터에 의해서 메모리가 낭비된다는 사실!
    - 기본 Data Parallel을 사용할 때는 하나의 디바이스가 데이터를 쪼개어 나누고, 각 디바이스에서의 처리 결과를 다시 모아서 계산하기 때문에 하나의 디바이스가 다른 디바이스에 비해 메모리 사용량이 많아진다. (2번째 GPU 참고)

텍스트, 스크린샷, 폰트, 메뉴이(가) 표시된 사진

자동 생성된 설명

* + - Deep Speed Zero는 MS에서 만든 대규모 분산 딥러닝을 위한 새로운 메모리 최적화 기술로 분산 학습 과정에서 불필요한 메모리의 중복 및 모델 및 데이터 병렬 처리에 필요한 리소스를 크게 감소시킬 수 있으며 학습할 수 있는 파라미터의 수를 크게 증가 가능하다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* + - 가장 빠른 BERT 학습 시간 달성 : 1,024개의 NVIDIA V100 GPU로 44분 소요

텍스트, 라인, 그래프, 폰트이(가) 표시된 사진

자동 생성된 설명

* + - * 위 그림은 단일 V100 GPU에서 DeepSpeed ZeRO, NVIDIA BERT, HuggingFace BERT의 성능을 비교한 결과입니다.
      * DeepSpeed는 max length 128, 512인 시퀀스 길이에 대해서 각각 64, 53 teraflops(초당 272, 52 샘플을 처리)의 처리량을 달성하여 NVIDIA BERT에 비해 최대 28%, HuggingFace BERT에 비해 최대 62% 향상된 처리량을 보여줍니다.
      * 또한 메모리 부족 없이 최대 1.8배 더 큰 배치 사이즈를 사용할 수 있습니다.

**Facebook OPT 125m 모델 학습:**

학습 환경 : Single A100 on Colab

Step1:

텍스트, 폰트, 스크린샷, 도표이(가) 표시된 사진

자동 생성된 설명

학습시간: 3분 39초

Batch size: 16

Max sequence length: 512

Learning rate: 1e-3

Epoch: 1

LoRA 사용하지 않음 (LoRA 사용 옵션, 공부하면서 어떻게 바뀌는지 확인해볼 예정)

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

모델 경로: output/actor-models

Step2:

텍스트, 폰트, 스크린샷, 도표이(가) 표시된 사진

자동 생성된 설명

학습시간: 7분

Batch size: 16

Max sequence length: 512

Learning rate: 5e-5

Epoch: 1

LoRA 사용하지 않음 (LoRA 사용 옵션, 공부하면서 어떻게 바뀌는지 확인해볼 예정)

모델 경로: output/reward-models