Appendix: Gradient Accumulations

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Batch Size

- 큰 배치사이즈는 epoch 내의 forward & backward 횟수를 줄여주어 학습의 속도를 높여줌

- 또한 배치사이즈에 따라 모델의 성능이 바뀔 수 있음
  - 작은 배치사이즈는 local minima를 탈출 할 수 있다고 알려져 있으나,
  - 큰 데이터셋에서는 배치사이즈가 클수록 오히려 성능이 높아지기도 함

- 따라서 모델의 성능이 떨어지지 않는 한도 내에서, 배치사이즈를 최대로 하여 학습을 빠르게 진행할 수 있음
  - 기본적으로 SGD를 사용할 경우, LR와 배치사이즈는 비례 관계를 갖게 됨
  - 하지만 Adam을 사용할 경우, LR에 크게 신경 쓸 필요 없음

- 하지만 GPU 메모리가 허락하지 않는다.

Gradient Accumulation

- Forward & backward를 할 때마다 파라미터 업데이트(optimizer.step() 호출)를 하는 대신, gradient를 누적해서 나중에 한번에 업데이트하는 방법
  - 마치 누적 횟수 만큼의 배치사이즈가 증가된 효과
  - e.g. k번 accumulation = $k \times$batch_size

$$\theta \leftarrow \theta + \nabla_\theta \sum_{i=1}^{N} y_i \cdot \log f_\theta(x_i)$$

- 속도 상의 이점은 없음

- Seq2seq에선 Adam optimizer 기준, batch_size 256이 가장 좋은 성능
  - --iteration_per_update 파라미터로 조절

How to Implement Gradient Accumulation