Setting Batch Size in Contrastive Learning

22.12.13. 이승한 (2020324009)

[방안 1] pre-trained NN to find K

2 stage approach

- step 1) **find # of modality (= M)**
  - let modality = number of clusters
  - using previous/proposed **clustering algorithm** (= pretrained)
- step 2) $\alpha \cdot M$ 을 batch size로

[방안 1] pre-trained NN to find K

2 stage approach

- step 1) **find # of modality (= M)**
  - let modality = number of clusters
  - using previous/proposed **clustering algorithm** (= pretrained)
- step 2) $\alpha \cdot M$ 을 batch size로

[방안 2] stratified sampling

2 stage approach

- step 1)
  - use previous/proposed **clustering NN algorithm** (= pretrained)
  - ex) 전체 데이터 1000개 : cluster 1,2,3,4,5 = (200, 100, 500, 50, 150)
- step 2)
  - stratified sampling

[방안 3] 일정 epoch 별 Data Loader 변경 / 누적 횟수 변경

(1 stage approach)

학습 중간 중간에, 일정 기준을 가지고 ..

- (1) data loader를 교체하는 방안은?
  (즉, epoch별로 적절한 data loader 사용)
- (2) gradient 누적횟수 변경

(1) data loader를 교체하는 방안

- ex) data loader # 1 : bs = 128
- ex) data loader # 2 : bs = 256
- ex) data loader # 3 : bs = 512

[방안 3] 일정 epoch 별 Data Loader 변경 / 누적 횟수 변경

(1 stage approach)

학습 중간 중간에, 일정 기준을 가지고 ..

- (1) data loader를 교체하는 방안은?
  (즉, epoch별로 적절한 data loader 사용)
- (2) gradient 누적횟수 변경

[방안 3] 일정 epoch 별 Data Loader 변경 / 누적 횟수 변경

(1 stage approach)

학습 중간 중간에, 일정 기준을 가지고 ..

- (1) data loader를 교체하는 방안은?
  (즉, epoch별로 적절한 data loader 사용)
- (2) gradient 누적횟수 변경

(1) data loader를 교체하는 방안

- ex) data loader # 1 : bs = 128
- ex) data loader # 2 : bs = 256
- ex) data loader # 3 : bs = 512

적절한 data loader의 판단 기준은? example

(epoch / iter 별 loss)

- epoch 1) data loader # 1 사용 시
  - iter 100 : ***0.75***
  - ...
  - iter 800 : ***0.73***
- epoch 2) data loader # 2 사용 시
  - iter 100 : ***0.65***
  - ...
  - iter 400 : ***0.60***
- epoch 3) data loader # 3 사용 시
  - iter 100 : ***0.59***
  - iter 200 : ***0.585***
- epoch 4) ...

[방안 3] 일정 epoch 별 Data Loader 변경 / 누적 횟수 변경

(1 stage approach)

학습 중간 중간에, 일정 기준을 가지고 ..

- (1) data loader를 교체하는 방안은?
  (즉, epoch별로 적절한 data loader 사용)
- (2) gradient 누적횟수 변경

(1) data loader를 교체하는 방안

- ex) data loader # 1 : bs = 128
- ex) data loader # 2 : bs = 256
- ex) data loader # 3 : bs = 512

적절한 data loader의 판단 기준은? example

(epoch / iter 별 loss)

- epoch 1) data loader # 1 사용 시
  - iter 100 : ***0.75***
  - ...
  - iter 800 : ***0.73***
- epoch 2) data loader # 2 사용 시
  - iter 100 : ***0.65***
  - ...
  - iter 400 : ***0.60***
- epoch 3) data loader # 3 사용 시
  - iter 100 : ***0.59***
  - iter 200 : ***0.585***
- epoch 4) ...

즉, 서로 다른 bs를 가진 data loader 후보군 K개를 가지고,

초반에 에폭 별로 다양하게 시도를 해봄 (dl # 1, .. dl # K)

이를 통해, loss를 가장 줄일 것으로 예상되는 dl을 매 step마다 선택

- 다만, 그 계산 과정이 (= dl별 loss 감소 예상치를 계산하는 과정이) 복잡하면 안됨
  - ex) extrapolation

[방안 3] 일정 epoch 별 Data Loader 변경 / 누적 횟수 변경

(1 stage approach)

학습 중간 중간에, 일정 기준을 가지고 ..

- (1) data loader를 교체하는 방안은?
  (즉, epoch별로 적절한 data loader 사용)
- (2) gradient 누적횟수 변경

(1) data loader를 교체하는 방안

- ex) data loader # 1 : bs = 128
- ex) data loader # 2 : bs = 256
- ex) data loader # 3 : bs = 512

적절한 data loader의 판단 기준은? example

(epoch / iter 별 loss)

- epoch 1) data loader # 1 사용 시
  - iter 100 : ***0.75***
  - …
  - iter 800 : ***0.73***
- epoch 2) data loader # 2 사용 시
  - iter 100 : ***0.65***
  - …
  - iter 400 : ***0.60***
- epoch 3) data loader # 3 사용 시
  - iter 100 : ***0.59***
  - iter 200 : ***0.585***
- epoch 4) …

즉, 서로 다른 bs를 가진 data loader 후보군 K개를 가지고,

초반에 에폭 별로 다양하게 시도를 해봄 (dl # 1, .. dl # K)

이를 통해, loss를 가장 줄일 것으로 예상되는 dl을 매 step마다 선택

- 다만, 그 계산 과정이 (= dl별 loss 감소 예상치를 계산하는 과정이) 복잡하면 안됨
  - ex) extrapolation

- ***loss값***들에대한 **extrapolation**을 해본 뒤, data loader를 교체한 것이 예상했던 값보다 (+) 인지, (-)인지에 따라!

[방안 4] gradient 누적 횟수/가중치 다르게

"몇 번"을 누적한 뒤 update를 진행할지 (=M), 그 값 자체를

들어오는 매 **input batch의 statistics에 dependent하게!**

- update가 이루어지기 위한 최소 요구치 = α
- α의 직관적 의미 : FULL dataset의 general information전부를 담고 있는 정도
- batch size를 크지 않게 시작함.
- batch 별 score : Distance (distn(full data), distn(batch data))
- ex)

  [epoch 1]

  - batch 1 (= 128) : 0.2α grad1
  - batch 2 (= 128) : 0.25α grad2
  - batch 3 (= 128) : 0.2α grad3
  - batch 4 (= 128) : 0.35α grad4
    - UPDATE (with 4번 누적)

  [epoch 2]

  - batch 1 (= 128) : 0.3α grad1
  - batch 2 (= 128) : 0.35α grad2
  - batch 3 (= 128) : 0.3α grad3
    - UPDATE (with 3번 누적)
  - ...

[방안 4] gradient 누적 횟수/가중치 다르게

"몇 번"을 누적한 뒤 update를 진행할지 (=M), 그 값 자체를

들어오는 매 **input batch의 statistics에 dependent하게!**

- update가 이루어지기 위한 최소 요구치 = α
- α의 직관적 의미 : FULL dataset의 general information전부를 담고 있는 정도
- batch size를 크지 않게 시작함.
- batch 별 score : Distance (distn(full data), distn(batch data))
- ex)

  [epoch 1]

  - batch 1 (= 128) : 0.2α grad1
  - batch 2 (= 128) : 0.25α grad2
  - batch 3 (= 128) : 0.2α grad3
  - batch 4 (= 128) : 0.35α grad4
    - UPDATE (with 4번 누적)

  [epoch 2]

  - batch 1 (= 128) : 0.3α grad1
  - batch 2 (= 128) : 0.35α grad2
  - batch 3 (= 128) : 0.3α grad3
    - UPDATE (with 3번 누적)
  - ...

[epoch 1]의 update

grad 1 ~ grad4를 누적해서 :

- (0.2*grad1) + (0.25*grad2) + (0.2*grad3) + (0.35*grad4) * Normalizing Constant
  - Normalizing Constant = 1/(0.2+0.25+0.2+0.35)