cleanlab(ICML Workshop 2023)

초록.-----

문제 : semantic segmentation 분야에서 라벨링 과정은 픽셀마다 사람이 라벨링 해줘야 해서 에러가 있기 쉽다. 그래서 gt 만들기가 어렵다. 이런 noisy label 로 모델을 학습하고 평가하는 것은 굉장한 문제이다.

Ours : noisy label 을 자동으로 감지하는 알고리즘 연구 -> 즉, 라벨 품질 평가하는 알고리즘 연구
(noisy label 이다 == 라벨품질이 낮다)
이때 이 label quality 는 학습된 segmentation 모델의 예측 값 기반해서 점수 매김.

Contribution : 라벨 품질 평가하는 7가지 방법. (Ours 는 Softmin 방법)

1.-----

noisy label 유형 3가지
1. Drop : 라벨링 안함 이슈
2. Swap : 잘못된 라벨링 이슈
3. Shift : 마스크 이슈

주의. 본 논문에서는 'truth' 라는 용어를 'noise 아예 없는 완벽한 이상적인 gt'라는 의미로 새로 정의.(gt 에는 noise가 있을수 있기 때문에 구분하기 위해)

cleanlab(ICML Workshop 2023)

2.-----

3.-----

결국 목표는, 이미지 x 에 대하여 라벨 품질 점수 s(x) 얻는것.

그리고 ours 는 softmin 방법 제안.

3.1. CCP

모델 예측 결과와 라벨이 일치하는 비율

$$s_{CCP}(x) = \frac{\sum_{i,j} \mathbb{I}[l_{ij} = P_{ij}]}{h \cdot w} \tag{1}$$

3.2. TCCP

CCP 에 클래스 별 임계값 도입,

그리고 마지막에 평균

$$s^k_{TCCP,t}(x) = \frac{\sum_{i,j} \mathbb{I}[l_{ij} = k, p_{ijk} > \tau]}{h \cdot w} \tag{2}$$

(3) 식 어떻게 구한다는건지 다시 봐야할듯.

$$\tau^*_k = \underset{\tau \in T}{\operatorname{argmax}}\, s^k_{TCCP,\tau}(x) \tag{3}$$

$$s_{TCCP}(x) = \frac{1}{K} \sum_{k=1}^{K} s^k_{TCCP,\tau^*_k}(x) \tag{4}$$

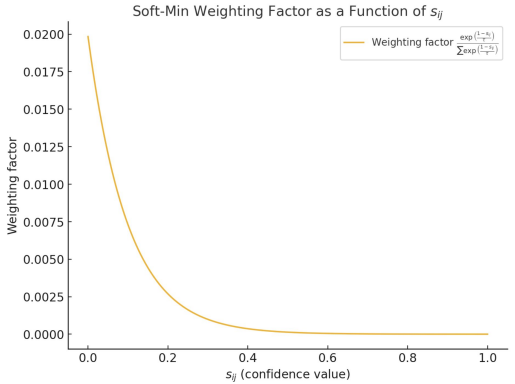
cleanlab(ICML Workshop 2023)

3.3. CIL

픽셀 별 라벨 품질 점수 구하고 이미지 평균

$$s_{CIL}(x) = \frac{1}{h \cdot w} \sum_{i,j} s_{ij} \tag{5}$$

$$s_{ij} := p_{i,j,l_{ij}}$$



3.4. Softmin(본 논문의 제안 방법)

CIL 의 한계 :

1) 라벨이 잘못되었을 경우 픽셀의 라벨품질점수가 낮을 것

2) 라벨이 정확하다 해도, 모델의 예측 결과 변동에 따라 확확 바뀜.

점수가 제일 낮은 픽셀 = 제일 자신감 없는 픽셀

그러나, 점수가 제일 낮은 것만을 쓰면은 전반적인 이미지 점수를 무시하는 거여서 바람직하지 않음.

그래서 대신에, 점수의 최소값을 soft approximation 한것을 사용.

$$s_{SM}(x) = \sum_{i,j} s_{ij} \cdot \frac{\exp\left(\frac{1-s_{ij}}{\tau}\right)}{\sum_{i,j} \exp\left(\frac{1-s_{ij}}{\tau}\right)} \tag{6}$$

cleanlab(ICML Workshop 2023)

3.5. CLC

Northcutt, C. G., Jiang, L., and Chuang, I. L. Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research, 70:1373–1411, 2021b.

이 논문에서 classification task 에서 LED 하던 것을 segmentation task 로 확장함

classification task 에서 (이미지와 무관하게) 각 픽셀을 독립적인 객체로 취급하고, confident learning 함, 이진 마스크 b 가 inference 아웃풋. (binary mask b 는 이미지 x 내에서 어떤 픽셀이 잘못 라벨링 되었는지 추정하는 마스크)

메트릭 설명 : 이미지 내 라벨링 잘된 픽셀 비율

$$s_{CLC} = \frac{\sum_{i,j} b_{ij}}{h \cdot w} \tag{7}$$

3.6. IOU

예측 마스크와 레이블 마스크 간 IoU 계산

$$s_{IOU}(x) = \frac{|\mathbf{P} \cap \mathbf{l}|}{|\mathbf{P} \cup \mathbf{l}|} \tag{8}$$

cleanlab(ICML Workshop 2023)

3.7. CoCo

Rottmann, M. and Reese, M. Automated detection of label errors in semantic segmentation datasets via deep learning and uncertainty quantification, 2022.

이 논문에서 LED 를 픽셀 수준에서 하는 것보다 연결_성분 수준에서 하는 것이 더 강건하다고 주장함.
이는 segmentation data 의 특징인, 이웃 픽셀들은 같은 클래스에 속할 확률이 높은 특징에 기반.

연결_성분 단위로 라벨품질점수가 매겨지고,
예측마스크던 라벨마스크던 모두, 공간적으로 인접한 픽셀들은 같은 클래스에 속함.

메트릭 구하는 과정.

1. 연결_성분 집합 만듬
2. 각 연결_성분이 해당 라벨에 속할 확률 구함 (픽셀마다의 확률 평균 내서)
3. 각 연결_성분의 라벨_품질_점수 구함
4. 각 이미지 내에 있는 연결_성분들의 평균 라벨_품질_점수 구함

$$p_c$$

$$s_c = p_c[k]$$

$$s_{CoCo}(x)$$

cleanlab(ICML Workshop 2023)

4. Experiments

4.1. 데이터셋 : SYNTHIA 데이터셋

noisy label 유형 3가지 <- 각 상황 만든 방법 아래 나열함.

1. Drop : 라벨링 안함 이슈 <- 선택된 클래스의 레이블을 랜덤하게 제거하여 해당 픽셀들을 unlabeled 카테고리에 넣음.

2. Swap : 잘못된 라벨링 이슈 <- 선택한 이미지 내에, 선택된 2개의 클래스에 해당되는 모든 레이블들을 랜덤하게 교환

3. Shift : 마스크 모양 (특히, 가장자리) 이슈 (주석자가 레이블 엉성하게 그린 상황) <- OpenCV 라이브러리 사용하여 세그맨테이션 마스크 모양에 변화 줌

벤치마킹 데이터셋으로 3개 사용.(각각 위 유형 중 하나씩 포함)

첫번째 데이터셋 : 20%가 Drop

두번째 데이터셋 : 30%가 Swap

세번째 데이터셋 : 20%가 Shift

학습 및 검증 이미지 개수 : 1112개, 1112개

검증 세트에 대해서만 예측 및 라벨_품질_점수 계산함.

4.2. 모델 : 1) DeepLabV3+, 2) FPN

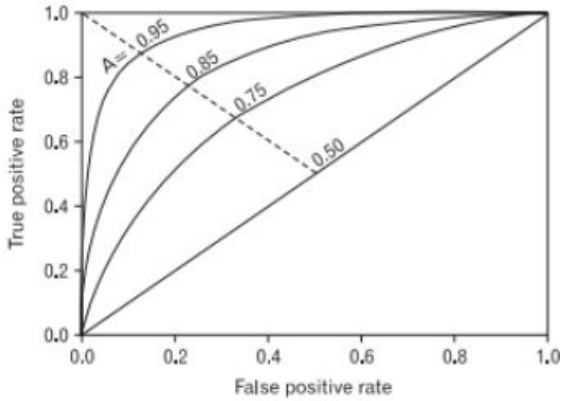
4.3. 평가

mislabeled 이미지 detect 하는 것 == information retrieval task -> 평가지표 : precision, recall

라벨_품질_점수가 mislabeled 이미지들을 얼마나 잘 랭킹 시키는지 평가 지표 : AUROC, AUPRC, Lift@T

1) AUROC : Area Under the Receiver Operating Characteristic Curve : ROC 곡선 아래의 면적

ROC 곡선 :



x축, y축 : FPR, TPR

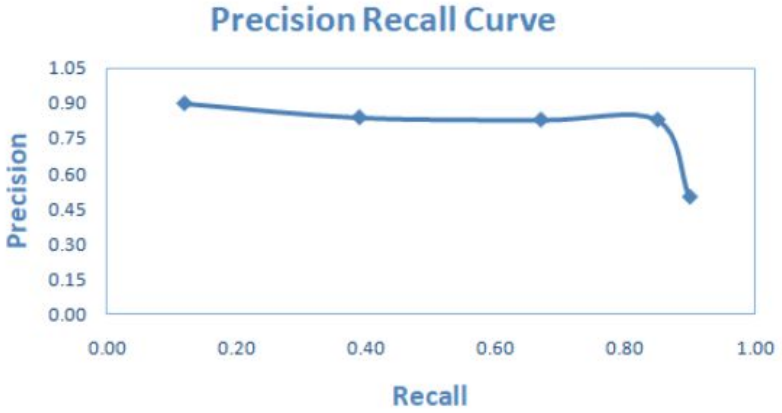
특징 : FP rate 와 TP rate 는 비례한다. TP rate 를 상승시키려면 FP rate 도 상승한다.

ROC 곡선의 면적이 1에 가까울수록 좋은 모델이다.

cleanlab(ICML Workshop 2023)

2) AUPRC : Area Under the Precision-Recall Curve : PRC 곡선 아래의 면적

PRC 곡선 :



x축, y축 : recall, precision

특징 : recall 과 precision 은 반비례 관계이므로, x축 오른쪽으로 갈수록 하향하는 곡선이다.

recall, precision 두 지표 모두 1에 가까울수록 좋은 모델이기 때문에, AUPRC 값도 1에 가까울수록 좋은 모델이다.

3) Lift@T : 전체 이미지 대비 상위 T개의 이미지에서 라벨 오류의 비율이 얼마나 높은지 평가

AUROC, AUPRC : precision, recall 둘다 평가 vs Lift@T : precision 만 평가

TP 가 드문 경우, AUPRC 가 AUROC 보다 더 informative 함.

1. AUPRC (Area Under Precision-Recall Curve)

- **정의**: Precision-Recall (정밀도-재현율) 곡선 아래의 면적.
  - **Precision (정밀도)**: $\frac{TP}{TP+FP}$
    모델이 "Positive"로 예측한 것 중 실제로 "Positive"인 비율.
  - **Recall (재현율)**: $\frac{TP}{TP+FN}$
    실제 "Positive"인 것 중 모델이 "Positive"로 잘 맞춘 비율.
- **해석**:
  - AUPRC 값이 높을수록 모델이 **"Positive 클래스"**를 잘 식별한다는 것을 의미.
  - 특히 데이터가 **불균형**한 경우 유용하다. (예: Positive 샘플이 적은 경우)
  - Positive 클래스의 **정확한 예측 능력**을 평가하는 데 중점을 둔다.
- **1에 가까울수록 좋은 이유**:
  - Precision과 Recall이 모두 높은 모델은 Positive 클래스를 정확하고 완벽하게 예측할 수 있음을 나타냄.

2. AUROC (Area Under Receiver Operating Characteristic Curve)

- **정의**: Receiver Operating Characteristic (ROC) 곡선 아래의 면적.
  - ROC 곡선은 **True Positive Rate (TPR)**와 **False Positive Rate (FPR)** 간의 관계를 나타냄.
    - **TPR (재현율)**: $\frac{TP}{TP+FN}$
      실제 Positive 샘플 중 잘 맞춘 비율.
    - **FPR**: $\frac{FP}{FP+TN}$
      실제 Negative 샘플 중 잘못 예측한 비율.
- **해석**:
  - AUROC 값이 높을수록 모델이 Positive와 Negative를 더 잘 구분함.
  - AUROC = 0.5는 무작위 추측과 같음을 의미. (성능 없음)
    AUROC = 1은 완벽한 분류기.
- **1에 가까울수록 좋은 이유**:
  - Positive와 Negative 클래스를 명확히 구분할 수 있다는 것을 의미.
  - 임계값에 관계없이 모델의 전반적인 분류 성능을 평가.