# Large Loss Matters in Weakly Supervised Multi-Label Classification

Youngwook Kim [1*], Jae Myung Kim [2*], Zeynep Akata [2,3], Jungwoo Lee [1,4]

[1]Seoul National University, [2]University of Tübingen, [3]Max Plank Institute, [4]HodooAI Lab

## Introduction

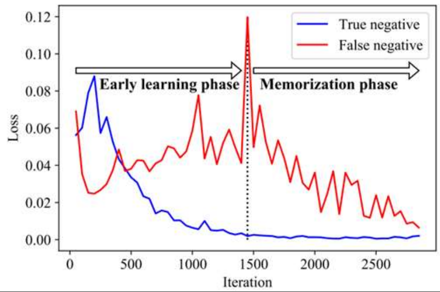
- Annotation cost for multi-label classification task is expensive.
  => Solution : annotate as "partial label"
- How to deal with unobserved labels?
  1. Explore cue of unobserved labels using bootstrapping, etc
     => Limitation : Requires heavy computation & optimization
  2. Assume all unobserved labels are negative
     => Limitation : Produces label noise
- Our work starts from the second approach. (Assume negative)

| | [a] | [b] |
|---|---|---|
| car | ✓ | ✓ |
| person | ✓ | |
| boat | ✗ | |
| bear | ✗ | ✗ |
| apple | ✗ | |

[a] : full label , [b] : partial label

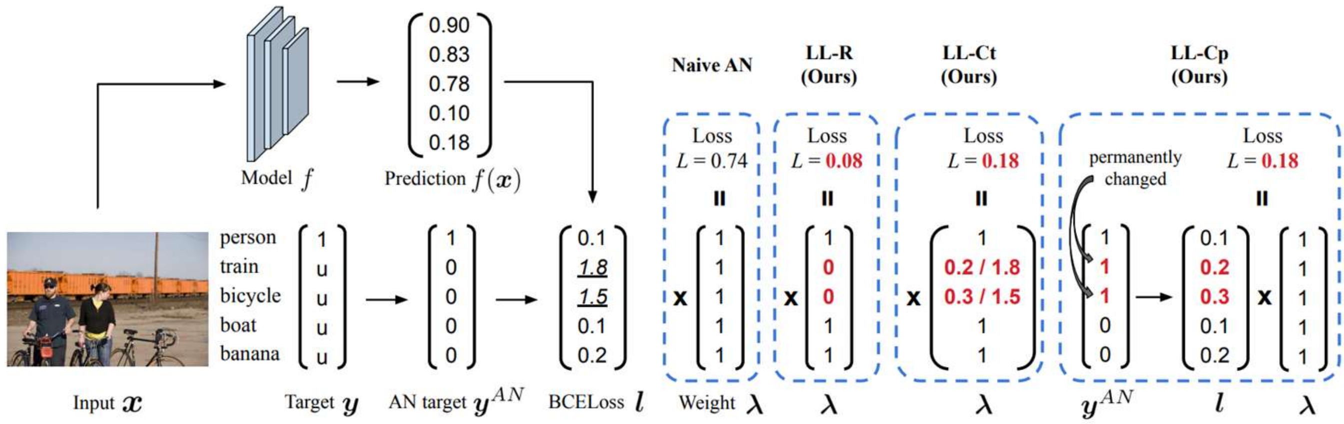
## Our Observation

- "Memorization effect" [1] occurs in a noisy multi-label setting.
- The model first fits into clean label
  and then gradually fits into noisy label.
  => Before memorization starts,
  noisy label usually results in a large loss during training. [2]
  => Large loss matters when training with partial label!




## Our Method

- Large loss rejection (LL-R) / temporal correction (LL-Ct) / permanent correction (LL-Cp)




## Our Results

**Artificially generated partial label datasets**

| Method | End-to-end | | | | LinearInit. | | | |
|---|---|---|---|---|---|---|---|---|
| | VOC | COCO | NUSWIDE | CUB | VOC | COCO | NUSWIDE | CUB |
| Full label | 90.2 | 78.0 | 54.5 | 32.9 | 91.1 | 77.2 | 54.9 | 34.0 |
| Naive AN | 85.1 | 64.1 | 42.0 | 19.1 | 86.9 | 68.7 | 47.6 | 20.9 |
| WAN [7,27] | 86.5 | 64.8 | 46.3 | 20.3 | 87.1 | 68.0 | 47.5 | 21.1 |
| LSAN [7,37] | 86.7 | 66.9 | 44.9 | 17.9 | 86.5 | 69.2 | 50.5 | 16.6 |
| EPR [7] | 85.5 | 63.3 | 46.0 | 20.0 | 84.9 | 66.8 | 48.1 | 21.2 |
| ROLE [7] | 87.9 | 66.3 | 43.1 | 15.0 | 88.2 | 69.0 | **51.0** | 16.8 |
| LL-R (Ours) | **89.2** | **71.0** | 47.4 | 19.5 | **89.4** | **71.9** | 49.1 | 21.5 |
| LL-Ct (Ours) | 89.0 | 70.5 | 48.0 | **20.4** | 89.3 | 71.6 | 49.6 | **21.8** |
| LL-Cp (Ours) | 88.4 | 70.7 | **48.3** | 20.1 | 88.3 | 71.0 | 49.4 | 21.4 |

**Real partial label dataset (OpenImages V3)**

| Method | G1 | G2 | G3 | G4 | G5 | All Gs |
|---|---|---|---|---|---|---|
| Naive IU | 69.5 | 70.3 | 74.8 | 79.2 | 85.5 | 75.9 |
| Curriculum [9] | 70.4 | 71.3 | 76.2 | 80.5 | 86.8 | 77.1 |
| IMCL [16] | 71.0 | 72.6 | 77.6 | 81.8 | 87.3 | 78.1 |
| Naive AN | 77.1 | 78.7 | 81.5 | 84.1 | 88.8 | 82.0 |
| WAN [7,27] | 71.8 | 72.8 | 76.3 | 79.7 | 84.7 | 77.0 |
| LSAN [7,37] | 68.4 | 69.3 | 73.7 | 77.9 | 85.6 | 75.0 |
| LL-R (Ours) | 77.4 | 79.1 | 82.0 | 84.5 | 89.5 | 82.5 |
| LL-Ct (Ours) | 77.7 | 79.3 | 82.1 | 84.7 | 89.4 | **82.6** |
| LL-Cp (Ours) | 77.6 | 79.1 | 81.9 | 84.6 | 89.4 | 82.5 |

## References

[1] Arpit et al., "A closer look at memorization in deep networks", ICML 2017

[2] Jiang et al., "MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels", ICML 2018