Neural Relation Graph: A Unified Framework for Identifying Label Noise and Outlier Data

Jang-Hyun Kim[1], Sangdoo Yun[2], Hyun Oh Song[1]

[1]Seoul National University [2]NAVER AI Lab

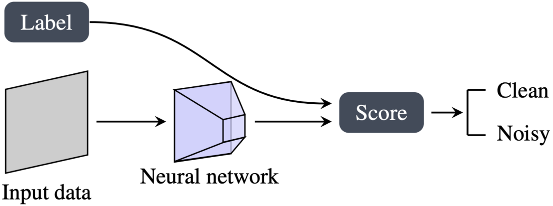
NeurIPS 2023

Goal of research

- **Dataset cleaning**: Identifying problematic data
  - Identifying problems regarding labels or input data
  - Developing domain-agnostic and scalable methods for label error and outlier detection

- **Data analysis**: Characterizing data points
  - Answering "Why does the model make such predictions?" from a data perspective
  - Building a more reliable evaluation system

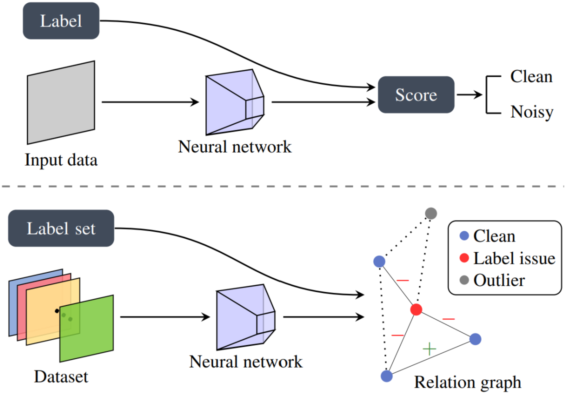
Conventional approach



- Conventional approach for identifying problematic data is to measure an **unary score** for each data:
  - prediction margin[1]
  - self-influence[2]
  - sensitivity[3]

[1] Northcutt et al., Confident learning: Estimating uncertainty in dataset labels, 2021
[2] Koh et al., Understanding black-box predictions via influence functions, 2017
[3] Liang et al., Enhancing the reliability of out-of-distribution image detection in neural networks, 2018

Proposed approach



- We propose a unified approach for detecting label noise and outlier data by utilizing **relational structure of data**.

Assumption

- Noisy training dataset $\mathcal{T} = \{(x_i, y_i) \mid i = 1, \ldots, n\}$.
  - May have problems in x_i (outlier) or y_i (label error).
- Trained neural networks on $\mathcal{T}$.
  - Extract feature representation $\mathbf{f}_i$.
  - Measure the **semantic similarity** $k : \mathcal{X} \times \mathcal{X} \to [0, M]$ between data (higher means more similarity).

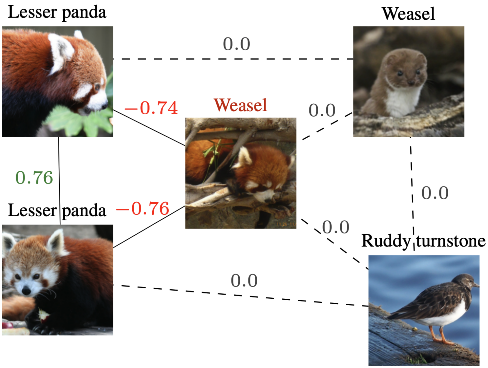
Data relation

- Given data (x_i, y_i) and (x_j, y_j), we define relation between data:

$$r\left((x_i, y_i), (x_j, y_j)\right) = 1(y_i = y_j) \cdot k(x_i, x_j).$$

  Here, $1(y_i = y_j) \in \{-1, 1\}$.

- Similar to the influence function, data relation quantifies the complementarity of a data pair.



Label error detection

- Goal: Measure the **label noisiness score** $s \in \mathbb{R}^n$ for dataset $\mathcal{T} = \{1, \ldots, n\}$.
  - A higher score indicates a higher likelihood of label error.

Label error detection

- Goal: Measure the **label noisiness score** $s \in \mathbb{R}^n$ for dataset $\mathcal{T} = \{1, \ldots, n\}$.
  - A higher score indicates a higher likelihood of label error.

- We consider a fully-connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$.
  - Node set $\mathcal{V} = \mathcal{T}$.
  - Weights $\mathcal{W}$ on edges $\mathcal{E}$ are the negative relation values:

$$w(i,j) = -r(i,j) = -r((x_i, y_i), (x_j, y_j)).$$

Label error detection

- Simple approach: Aggregate edge weights as $s[i] = \sum_{j=1}^{n} w(i,j)$.
  $\Rightarrow$ Edge values can affect both the clean and unclean data.

Label error detection

- Simple approach: Aggregate edge weights as $s[i] = \sum_{j=1}^{n} w(i,j)$.

  $\Rightarrow$ Edge values can affect both the clean and unclean data.

- We jointly estimate the **noisy subset** $\mathcal{N} \subset \mathcal{T}$ that are likely to have incorrect labels:

$$\mathcal{N}^* = \underset{\mathcal{N} \subset \mathcal{T}}{\operatorname{argmax}} \ \operatorname{cut}(\mathcal{N}, \mathcal{T} \setminus \mathcal{N}) \Big(\coloneqq \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{T} \setminus \mathcal{N}} w(i,j) \Big) - \lambda |\mathcal{N}|.$$

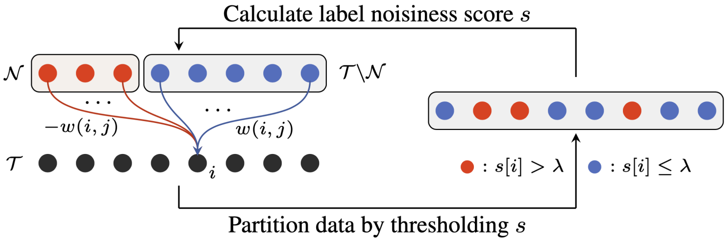
  $\Rightarrow$ Max-cut problem, which is NP-hard.

Label error detection

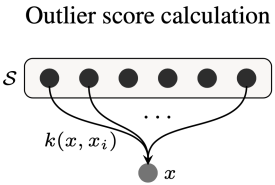
- Motivated by Kerninghan-Lin algorithm, we alternatively update s and $\mathcal{N}$:

$$s[i] = \sum_{j \in \mathcal{T} \setminus \mathcal{N}} w(i,j) - \sum_{j \in \mathcal{N}} w(i,j)$$

$$\mathcal{N} = \{i \mid s[i] > \lambda, i \in [1, \ldots, n]\}.$$



OOD/outlier detection



- We measure the **outlier score** (higher scores indicate greater outlierness) of a data point x as

$$\mathrm{outlier}(x) = \frac{1}{\sum_{i \in \mathcal{S}} k(x, x_i)}.$$

- Here, $\mathcal{S}$ is a random subset of $\mathcal{T}$.
  - Reflect global characteristics of data distribution.
  - Only 1% is enough in the case of ImageNet.

Kernel function

- We propose the following class of bounded kernel:

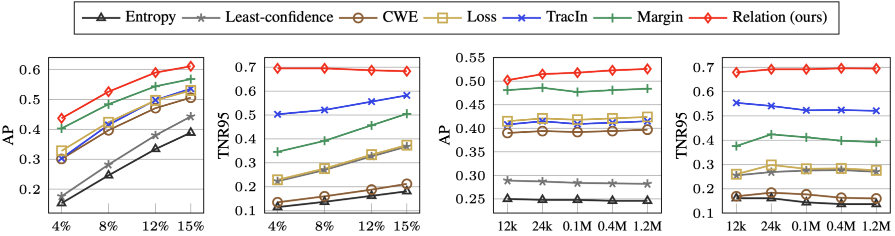
$$k(x_i, x_j) = |s(\mathbf{f}_i, \mathbf{f}_j) \cdot c(\mathbf{p}_i, \mathbf{p}_j)|^t,$$

  where hyperparameter $t > 0$ controls the kernel distribution's sharpness.
  - Feature similarity: $s(\mathbf{f}_i, \mathbf{f}_j) = \max(0, \cos(\mathbf{f}_i, \mathbf{f}_j))$
  - Prediction compatibility: $c(\mathbf{p}_i, \mathbf{p}_j) = P(\widehat{y}_i = \widehat{y}_j) = \mathbf{p}_i^\intercal \mathbf{p}_j$

- Our framework demonstrates strong performance across various kernel types, including RBF kernels.

Experiment results: Label error detection

- An MAE-Large model on ImageNet with synthetic label noise.



(a) Label noise ratio (%)

(b) The number of data (log-scale)

Experiment results: Label error detection

- Detected data samples with label errors from ImageNet and SST2 (text sentiment classification).

Negative

"entertaining and informative documentary"

Positive (-0.850)

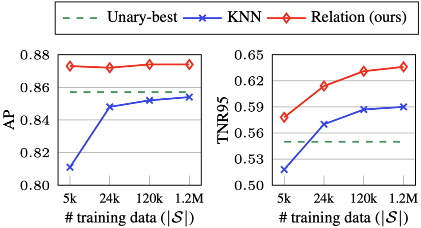
"entertaining movie"

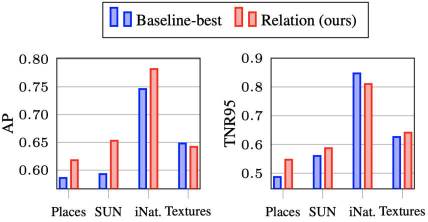
Positive (-0.846)

"fascinating and timely content"

Experiment results: OOD detection

- An MAE-Large model on ImageNet validation set with various OOD datasets.





Experiment results: Outliers in validation set

- Detected outlier validation samples from ImageNet (top) and SST2 (bottom).

Flute | Airship | Vase | Honeycomb | Maze | Alligator lizard | Cockroach

Positive

“leather pants"

Positive

“give a backbone to the company"

Negative

“the israeli/palestinian conflict as"

Summary

- We propose a unified approach for identifying label errors and outlier data points.
- We develop domain-agnostic and scalable detection algorithms.
- https://github.com/snu-mllab/Neural-Relation-Graph