Deep Image Retrieval is not Robust To Label Noise

Stanislav Dereka, Ivan Karpukhin, Sergey Kolesnikov @ Tinkoff AI

TINKOFF

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Label noise in Image Retrieval (IR)

- Large-scale datasets are essential for the success of deep learning in image retrieval. However, manual assessment errors and semi-supervised annotation techniques can lead to label noise even in popular datasets. It was shown recently that the annotation error rate in large-scale datasets can exceed 40%. Previous work show that image classification is to some extent robust to label noise: deep classification models can be trained on partially mislabelled data without a significant performance drop.
- There are multiple reasons which make label noise affect differently IR and classification tasks: (1) IR datasets typically include many (thousands) of classes; (2) these classes are imbalanced; (3) IR problems are open-set, i.e. train and test classes don't intersect.
- In this work we (1) show that IR methods are less robust to label noise than image classification ones. Furthermore, we (2) investigate different IR-specific label noise types and study their effect on model's performance.

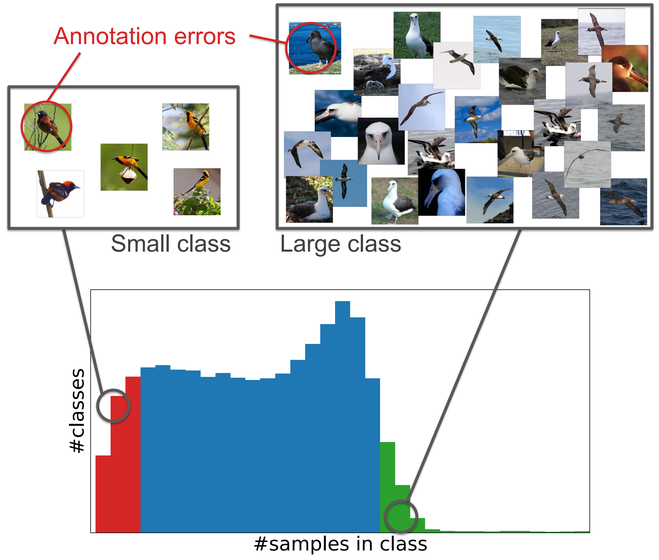

Fig 1. IR datasets typically contain thousands of categories which are severely imbalanced.

IR-specific label noise types

- As shown in Fig. 1, image retrieval datasets are highly class-imbalanced. By selecting a fixed amount of noisy items, we can choose either many small classes or a small number of large classes when we model label noise.
- In **large class label noise** we select classes from the right tail of the distribution in Fig. 1 for assigning wrong labels, in **small class label noise** — from the left tail.
- In **uniform label noise** a given fraction of samples is selected independently of their class labels.

IR and classification label noise robustness

- We vary the proportion of mislabeled samples in training dataset and measure relative performance drop (compared with performance on clean dataset) for IR and classification tasks.

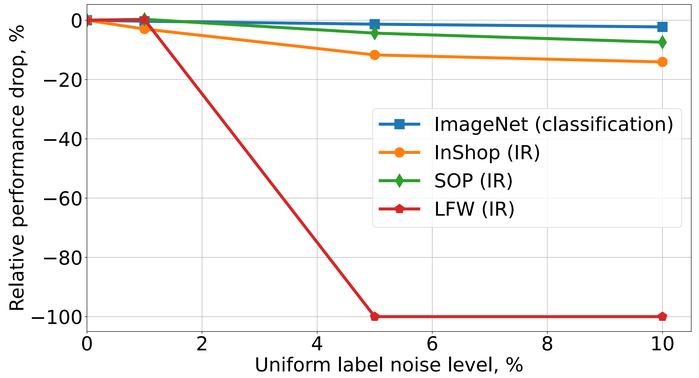

Fig 2. Connection between noise level and performance drop. We measure R@1 for IR and accuracy for classification.

- It can be seen (Fig. 2) that IR tasks are times more sensitive to annotation errors in train dataset than image classification ones. Moreover, in some tasks (LFW) **even 10% of mislabeled samples can completely ruin model's performance.**

Comparing label noise types' effect

- It can be seen from Fig. 3 that **the more training set classes are affected by label noise, the higher is the drop in performance**, even if the total amount of corrupted elements doesn't change.
- Small class label noise leads to more significant degradation than large class label noise. The effect of uniform label noise is the strongest, as it affects more classes than the other noise types. We can also conclude, that **annotation errors in small classes have the strongest effect on performance drop.**

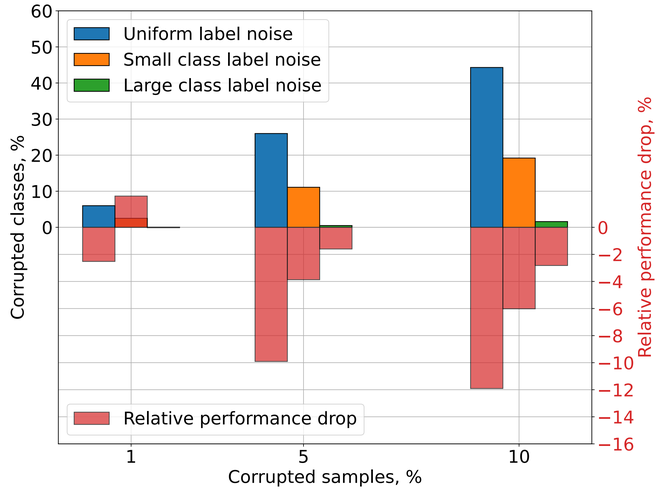

Fig 3. Connection between corrupted classes, corrupted samples proportions, and relative quality drop for InShop dataset.

Authors' contacts:
st.dereka@gmail.com
i.a.karpukhin@tinkoff.ru
scitator@gmail.com

Full paper: