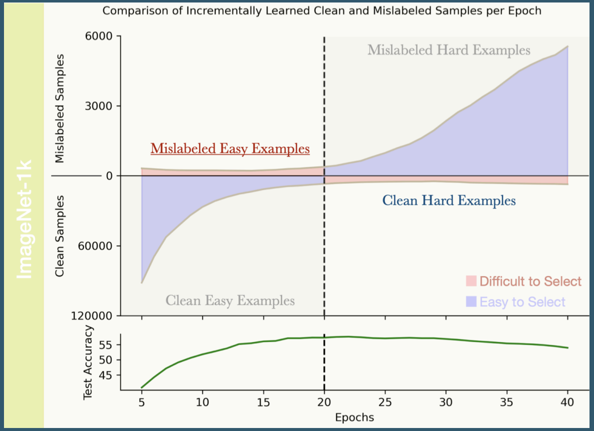
Enhancing Sample Selection Against Label Noise by Cutting Mislabeled Easy Examples

NEURAL INFORMATION PROCESSING SYSTEMS

Suqin Yuan, The University of Sydney
Lei Feng, Southeast University
Bo Han, Hong Kong Baptist University
Tongliang Liu, The University of Sydney



1. Challenges in Sample Selection

- **Sample Selection**: identify a "confident" subset of samples (presumed to be clean) and use them for (re-)training.
- It's well-known that DNNs learn simple, clean patterns first, before gradually Memorization. This leads to a common assumption: Samples learned early in training / samples the model masters well, are clean and trustworthy.

The Challenges:

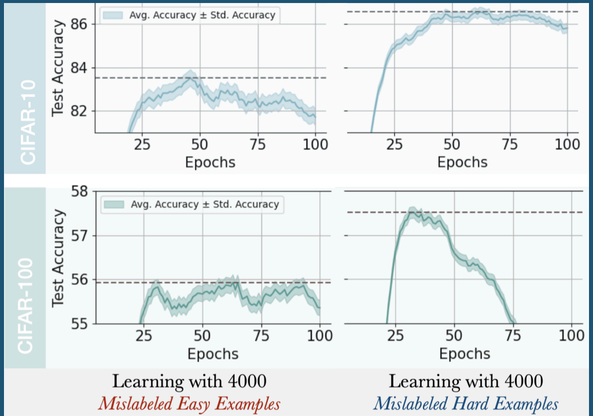
1. ~~Removed samples may contain clean samples.~~ (and they are often important!!! See our ICCV'21/23, NeurIPS'23/24)

2. Selected samples may actually be mislabeled (Our research object here!!!).

2. Early Learned Mislabeled Samples (MEEs) are Most Harmful

Mislabeled Easy Examples (MEEs)

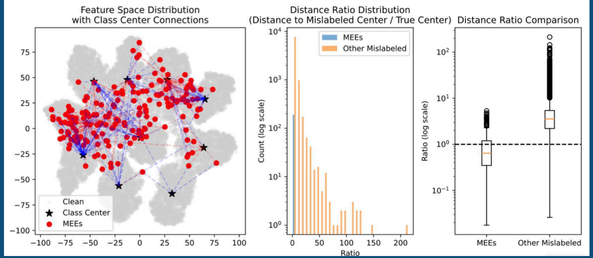
- We define a new subset of noisy data: **MEEs.** These are mislabeled samples that the model *incorrectly* but *confidently* learns (correct classification) early in the training process.
- We category all mislabeled samples into 5 groups based on when the model learned them (from earliest (MEEs) to latest (MHardEs)). We then trained new models using the clean data combined with only one of these mislabeled groups.



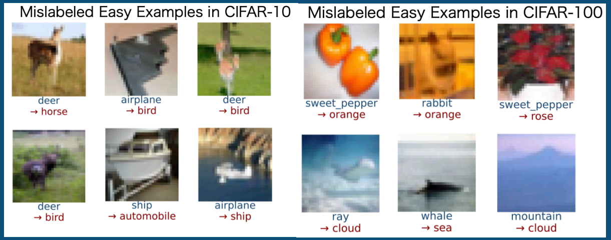
The mislabeled samples learned earliest are the most harmful to model generalization. Relying on early-learning dynamics to select "clean" data is a flawed strategy. (Very few in number, but harmful)

3. Deep Understanding: MEEs

- MEEs are not random errors. They are "reasonably" mislabeled, meaning their features create a strong, misleading signal for the model.



- MEEs is closer to incorrect wrong labels (class center) than their true labels (class center) in the learned feature space, at very beginning of training.



- MEEs are learned early because their misleading features perfectly align with the simple, low-level patterns (like color or texture) that DNNs extract in the initial training stages. This poisons the model's feature representations from the very beginning, leading to poor generalization.