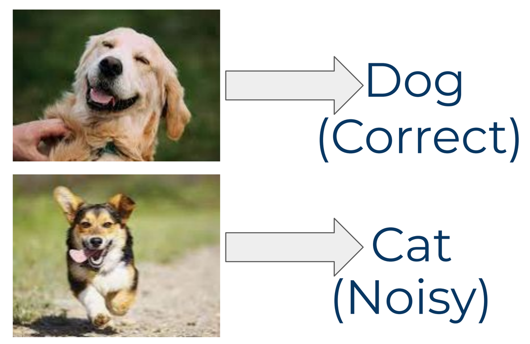
# A Framework using Contrastive Learning for Classification with Noisy Labels

M. Ciortan, R. Dupuis, and T. Peels

## 1. Introduction

**Context: Noisy-label learning**




Acquiring high-quality human labels for classification can be expensive or time consuming. Cheaper alternatives (e.g. web crawlers) can lead to samples with noisy labels.

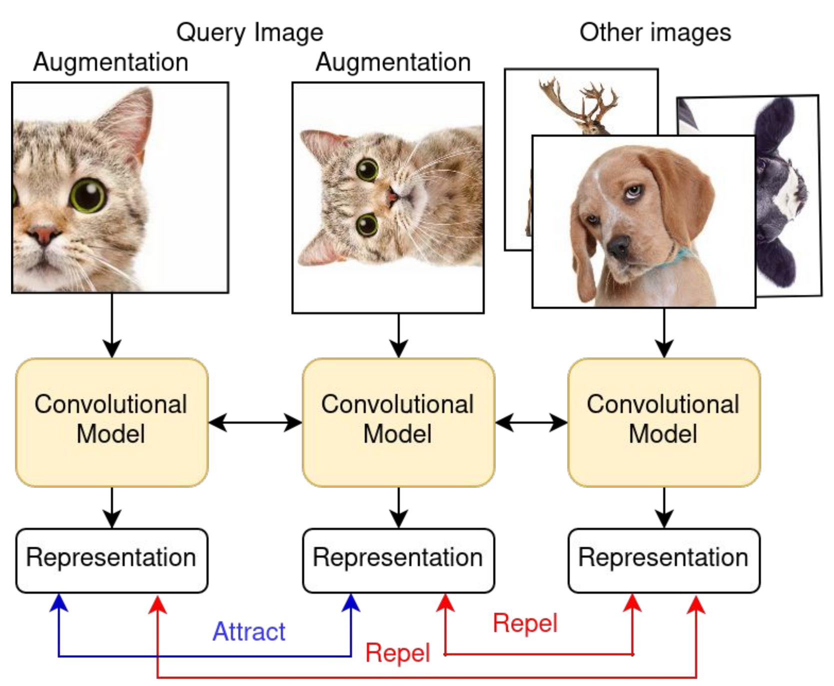
**Problems**

Deep neural networks tend to overfit to noisy labels due to their high capacity. Moreover, performing classical early-stopping to avoid such overfitting is very difficult in the absence of a clean validation set.

**Proposed solutions**

- Reducing the overfitting to noisy labels by pre-training representations to be transferred to the classifier trained on noisy labels by fine-tuning.
- Using robust loss functions less sensitive to noisy labels.
- Filter label noise by splitting the per-sample loss distribution into two groups: clean and noisy data (noisy data have a larger loss).

## 2. Contrastive learning and robust losses




**Contrastive learning**

Self-supervised contrastive learning algorithms have made great progress in extracting features (Moco, SwAV, etc). SimCLR[1] is used in our framework.

The central idea is to bring different instances of the same input image (data augmentation) closer and spread instances from other images apart. The inputs are divided into positive and negative pairs.

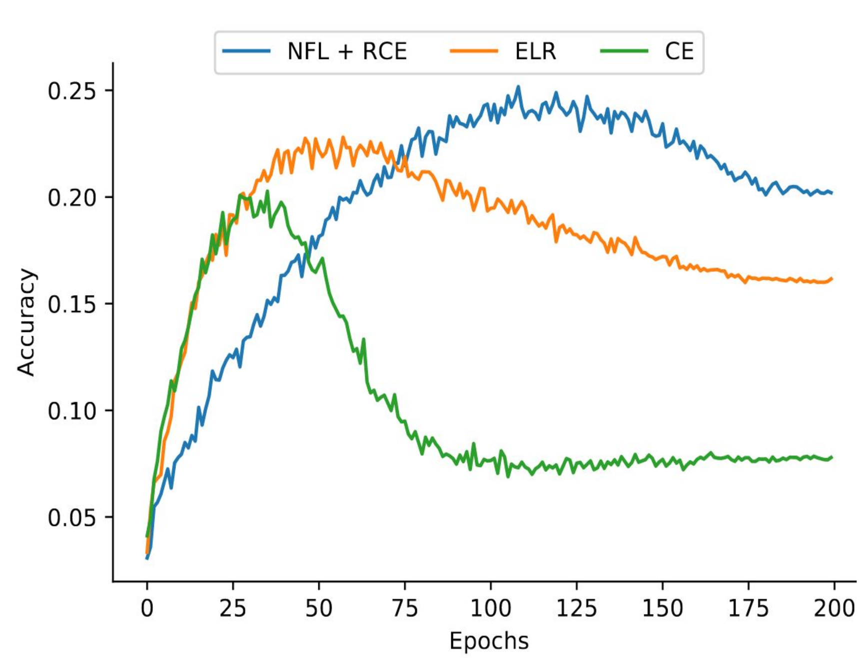
**Robust loss functions**

Various loss functions that are inherently noise-tolerant have been developed. Theoretical and empirical results show that such loss functions increase the accuracy of the classifier under noisy labels.

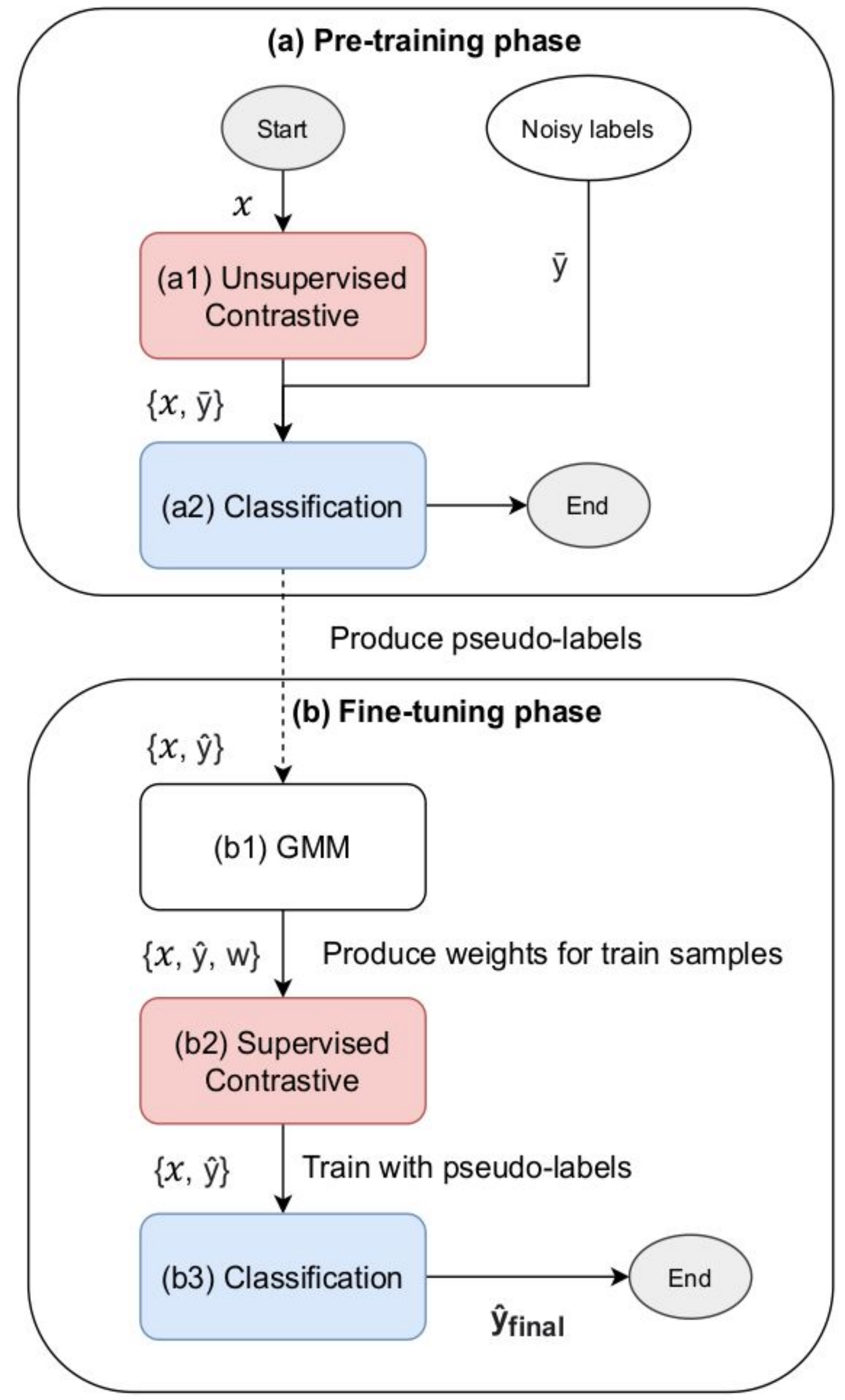
Three losses are used in this work:

- Cross Entropy (CE)
- Normalized Loss Function (NFL)[2] with Reverse Cross Entropy (RCE)[2]
- Early Learning Regularization (ELR)[3]

The last two are robust. However, as shown in the Figure, even if they perform better, some overfitting remains.




## 3. Proposed framework




The framework is composed of three parts: two main blocs and a pseudo-label generation between the two.

**Pre-training bloc (a)**

Representations are computed by contrastive learning and transferred to a classifier trained on noisy labels with one of the three losses (CE, NFL+RCE, and ELR). It could also be substituted by more state-of-the-art methods such as DivideMix[4] or ELR+[3].

**Pseudo-labels generation:**

Based on the assumption that the training set labels, predicted after training the model with a pre-training and a noise-robust loss function are more accurate than the ground truth.

**FIne-tuning phase (b)**

- Sample selection: It leverages the pseudo-labels predicted by the bloc (a). The label noise is filtered with the small loss trick. It identifies clean and noisy samples by considering a certain number of small-loss training samples as true-labeled. The sample correctness probability is computed by fitting a 2 components Gaussian Mixture Model (GMM).

- Weighted supervised contrastive learning: To avoid confirmation bias, the pseudo-labels are first used to compute updated representations.

$$L_i = -log \frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\widetilde{w_{p,i}} exp(z_i^T z_p / \tau)}{\sum_{a \in A(i)} exp(z_i^T z_p / \tau)}$$

Weighted supervised formulation

$z$ is a feature vector, $P(i)$ is the set of batch samples belonging to the class i, and $A(i)$ is the set of negative samples. $\tau$ is a temperature.

This loss is supervised[5]: the set of all samples from the same class is considered positive (the unsupervised version considers only augmentations) against the negatives from the remainder of the batch.

The sample correctness probability weights the supervised contrastive learning loss with a modified weight $\widetilde{w_{p,i}}$: equal to 1 if i=p and equal to the sample correctness probability otherwise.

## 4. Results

**Datasets**

- Cifar10/100: 60.000 32x32 images in 10 or 100 classes. The dataset is contaminated with simulated symmetric or asymmetric label noise.
- Webvision and Clothing 1M: Real-world dataset with unknown noise ratios. Images are of higher dimensions.

**Gains from the contrastive pre-training**

We compare baseline classifier with the pre-trained version for the three losses on various noise levels. The pre-training outperforms the baselines by large margins for CIFAR10 (same for CIFAR100).

| | | | CIFAR10 | |
|---|---|---|---|---|
| Type | $\eta$ | Loss | Base | Pre-t. |
| Sym | 0.2 | ce | 77.2 | **87.7** |
| | | elr | 90.3 | **93.0** |
| | | nfl+rce | 91.0 | **92.7** |
| | 0.4 | ce | 58.2 | **78.0** |
| | | elr | 82.3 | **92.0** |
| | | nfl+rce | 87.0 | **91.4** |
| | 0.6 | ce | 35.2 | **59.2** |
| | | elr | 64.2 | **90.4** |
| | | nfl+rce | 80.2 | **88.1** |
| | 0.8 | ce | 17.0 | **27.3** |
| | | elr | 18.3 | **84.8** |
| | | nfl+rce | 42.8 | **59.9** |

**Impact of the fine-tuning**

The gain from the fine-tuning is more limited. The real-world datasets show an average accuracy improvements of 1.8%.

| | Webvision | | | Clothing1M | | |
|---|---|---|---|---|---|---|
| Loss | Base. | Pre-t. | Fine-Tune | Base. | Pre-t. | Fine-Tune |
| ce | 51.8 | 57.1 | **58.4** | 54.8 | 59.1 | **61.5** |
| elr | 53.0 | 58.1 | **59.0** | 57.4 | **60.8** | 60.4 |
| nfl+rce | 49.9 | 54.8 | **58.2** | 57.4 | 59.4 | **60.1** |

**Comparison with the state of the art**

The proposed framework does not outperform state-of-the-art methods such as DivideMix, ELR+, SELF. However, they are based on dual networks, MixUp, semi-supervised learning, or ensemble learning, etc. introducing extra hyperparameters, complexity, and difficulties in reproducibility.

## 5. Main findings and comments

- Contrastive pre-training boosts performance robustly and reduces the sensitivity to noisy labels but increases the computational cost.
- Robust loss functions are also improved by the pre-training.
- The framework is more stable to changes in the hyperparameters but it remains sensitive similar to competing methods of the SOTA.
- This paper addresses a timely topic as shown by numerous recent studies in 2020/21 on contrastive learning and noisy labels classification.

## References

[1] Chen et al., “A simple Framework for Contrastive Learning of VIsual Representations”, 2020
[2] Ma et al., “Normalized loss functions for deep learning with noisy labels”, ICML, 2020
[3] Liu et al., “Early-Learning Regularization Prevents Memorization of Noisy Labels”, NeurIPS, 2020
[4] Li et al., “DivideMix: Learning with Noisy labels as semi-supervised Learning”, ICLR, 2020
[5] Khosla et al., “Supervised Contrastive Learning”, NeurIPS, 2020