Robust Classification on Noisy labels

Raushan Kumar (200070068)
Dept. of Electrical Engineering
IIT Bombay

***Abstract*—Robustness to noisy labels is a crucial challenge in machine learning, as real-world datasets often contain annotation errors. Contrastive learning has emerged as a promising technique for addressing this issue. In this report, we explore two contrastive learning approaches for learning with noisy labels: "Learning to Learn from Noisy Labeled Data" and "Twin Contrastive Learning with Noisy Labels" (TCL). Our focus was implementing TCL, which proposes a novel framework integrating Gaussian Mixture Models, out-of-distribution detection, and cross-supervision to achieve robust learning in noisy environments. We discuss the theoretical foundations of TCL and detail our implementation efforts.**

While we encountered 'CUDA out of memory' issues while attempting to implement "Learning to Learn from Noisy Labeled Data," even with a batch size of 1, a thorough analysis revealed its core strengths. This approach tackles a noise-tolerant training algorithm, where a meta-learning update is performed before a conventional gradient update. The proposed meta-learning method simulates actual training by generating synthetic noisy labels and trains the model such that after one gradient update using each set of synthetic noisy labels, the model does not overfit to the specific noise.

***Index Terms*—Deep Learning, Noisy Labels, Contrastive Learning, Twin Contrastive Learning (TCL), Gaussian Mixture Models (GMMs), Out-of-Distribution (OOD) Detection, Cross-Supervision, CIFAR-10, CIFAR-100, Resource Constraints, Image Classification, Clothing, Entropy Regularization, Instance-wise Discrimination**

I. Introduction

Deep neural networks (DNNs) excel in image classification tasks but their success is highly dependent on large-scale, carefully curated datasets with accurate labels. Unfortunately, obtaining such datasets is costly and time-consuming. Real-world datasets often contain noisy labels, either due to less rigorous annotation or being sourced from platforms where accuracy isn't guaranteed (e.g., online marketplaces, crowdsourcing). Label noise poses a major hurdle for DNNs, as they tend to overfit to these errors, leading to significant performance degradation.

Researchers have sought various ways to mitigate the impact of noisy labels. Several approaches focus on designing robust loss functions, reducing the weights of potentially noisy samples, or attempting to correct the labels directly. These label correction methods, involving iterative sample selection processes, have shown promise. Recently, contrastive learning has emerged as a powerful technique for addressing noisy labels. Contrastive methods focus on learning discriminative representations, with some approaches later cleaning labels or creating positive sample pairs leveraging nearest neighbor information. However, nearest-neighbor techniques are susceptible to scenarios with extremely high noise, as neighboring samples themselves might be mislabeled.

This report investigates two contrastive learning-based approaches to combat the challenges posed by noisy labels:

Learning to Learn from Noisy Labeled Data (Paper 1): This paper introduces a meta-learning-based noise-tolerant training algorithm. The core idea is to simulate noisy labels during a meta-learning phase. By making the model resistant to overfitting on these synthetically noisy labels, the goal is to train it toward greater generalization and noise robustness on the real, noisy dataset.

Twin Contrastive Learning with Noisy Labels (TCL) (Paper 2): TCL offers a novel framework that combines the strengths of contrastive learning and Gaussian Mixture Models (GMMs) for noise-tolerant learning. First, it learns discriminative image representations in an unsupervised manner using contrastive learning. It then constructs a GMM over these representations, going beyond typical unsupervised GMMs by linking the latent variables to the model's predictions. TCL addresses the limitation of nearest-neighbor approaches by formulating a novel out-of-distribution (OOD) label noise detection method. It does this by using a two-component GMM to model samples with clean and wrong labels, thus considering the full data distribution.

II. Prior Work

Robustly handling noisy labels is a persistent challenge in machine learning. Contrastive learning focuses on learning discriminative representations. While traditional contrastive methods primarily target unsupervised representation learning.

Robust Loss Functions: These approaches modify loss functions to reduce the impact of noisy samples on model training [1], [2], [3], [4], [5], [6].

Noise Modeling: These methods seek to characterize label noise distributions, using neural networks, probabilistic models, or knowledge graphs to infer true labels [5]. However, they often require clean reference data or are limited by the assumptions of the noise model.

Label Correction: These iterative techniques attempt to identify and modify noisy labels during training, often leveraging consistency between data augmentations and model predictions.

Noise Detection: Contrastive techniques are employed in identifying noisy labels, often by analyzing nearest neighbors in the embedding space. However, such approaches can be vulnerable in settings with extremely high noise rates where nearby samples are also likely to be mislabeled.

Pre-training: Contrastive learning has been used as a pre-training step to improve model robustness [6], but its direct effectiveness in the presence of label noise remains limited.

III. Datasets Used

we have used CIFAR-10 dataset in our work to experiment with but the original paper conducted experiments on several datasets:

- Cifar-10
  - Collection of Images: Composed of 60,000 color images, each 32x32 pixels in size.
  - Classes: Images are divided into 10 mutually exclusive classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck
  - Training & Testing Splits: Training Set: 50,000 images (5,000 images per class)
  - Testing Set: 10,000 images (1,000 images per class)
- Cifar-100
  - Collection of Images: Comprises 60,000 color images, each 32x32 pixels in size, just like CIFAR-10.
  - Classes: The significant difference lies in the number of classes. CIFAR-100 has 100 different classes. These classes are grouped into 20 'superclasses' for an additional level of categorization.
  - Training & Testing Splits: **Training Set**: 50,000 images (500 images per class)
  - Testing Set: 10,000 images (100 images per class)
- Clothing1M
  - Scale: A large-scale dataset containing one million clothing images.
  - Noisy Labels: The key characteristic of the dataset is that the labels are collected from websites and are inherently noisy (i.e., often contain mislabeled examples).
  - Classes: 14 clothing categories (e.g., t-shirt, dress, jeans, etc.).
  - Clean Dataset Subsets: The dataset also includes smaller subsets with clean labels:
  - 50,000 images for training
  - 14,000 images for validation
  - 10,000 images for testing
- Webvision
  - Purpose: Designed to facilitate research on robust learning from noisy, real-world web data.
  - Composition: Contains over 2.4 million images collected from Flickr and Google Images.
  - ImageNet Connection: Intentionally uses the same 1,000 classes as the ImageNet ILSVRC 2012 dataset. This facilitates comparisons and investigations into transfer learning.
  - Web Noise: Images are inherently noisy, as they haven't been manually curated – labels might be incorrect or incomplete.

IV. Work Done

My work in BTP-2 began by studying the paper ”Learning to Learn from Noisy Labeled Data” and its references. I implemented the code from the associated repository. However, persistent ”CUDA OUT OF MEMORY” errors, even with minimal batch sizes, along with the paper's unpublished status, led me to abandon this approach. Subsequently, I focused on the paper ”Twin Contrastive Learning with Noisy Labels,” studying it in depth and implementing its corresponding code. Resource constraints on Google Colab and Kaggle presented challenges; I spent considerable time debugging and optimizing memory usage. Strategies included reduced batch sizes and saving only the most recent checkpoints. This paper implements 'Twin Contrastive learning with noisy labels' whose overview is shown in the figure.

Overview: TCL is a novel twin contrastive learning model designed for classification tasks where training data may contain noisy labels.It achieves this by leveraging contrastive learning to learn image representations and constructing a Gaussian Mixture Model (GMM) over these representations. Unlike unsupervised GMM, TCL links the label-free GMM and label-noisy annotations by replacing the latent variable of GMM with the model predictions for updating the parameters of the GMM.

This framework includes four main components:

- Modeling the data distribution via a GMM.
- Detecting examples with wrong labels as out-of-distribution samples.
- Cross-supervision by bootstrapping the true targets.
- Learning robust representations through contrastive learning and Mixup.

A. Modeling Data Distribution

Given the image dataset consisting of N images, we opt to model the distribution of x over its representation $v = f(x)$ via a spherical Gaussian mixture model (GMM). After introducing discrete latent variables $z \in \{1, 2, \ldots, K\}$ that determine the assignment of observations to mixture components, the unsupervised GMM can be defined as

$$p(v) = \sum_{k=1}^{K} p(v, z = k) = \sum_{k=1}^{K} p(z = k) N(v|\mu_k, \sigma_k). \quad (1)$$

where μ_k is the mean and σ_k a scalar deviation. If we assume that the latent variables z are uniform distributed, that is, $p(z = k) = 1/K$, we can define the posterior probability that assigns x_i to k-th cluster:

$$\gamma_{ik} = p(z_i = k|x_i) \propto N(x_i|\mu_k, \sigma_k). \quad (2)$$

To connect unsupervised GMM with available noisy labels y. Replace unsupervised $p(z_i = k|x_i)$ with noisy predictions $p_\theta(y_i = k|x_i)$.Link latent z with noisy labels y to guide GMM parameter updates:

$$\mu_k = \text{norm}\left(\frac{\sum_i p_\theta(y_i = k|x_i) v_i}{\sum_i p_\theta(y_i = k|x_i)}\right). \quad (3)$$

$$\sigma_k = \frac{\sum_i p_\theta(y_i = k|x_i)(v_i - \mu_k)(v_i - \mu_k)^T}{\sum_i p_\theta(y_i = k|x_i)}. \quad (4)$$

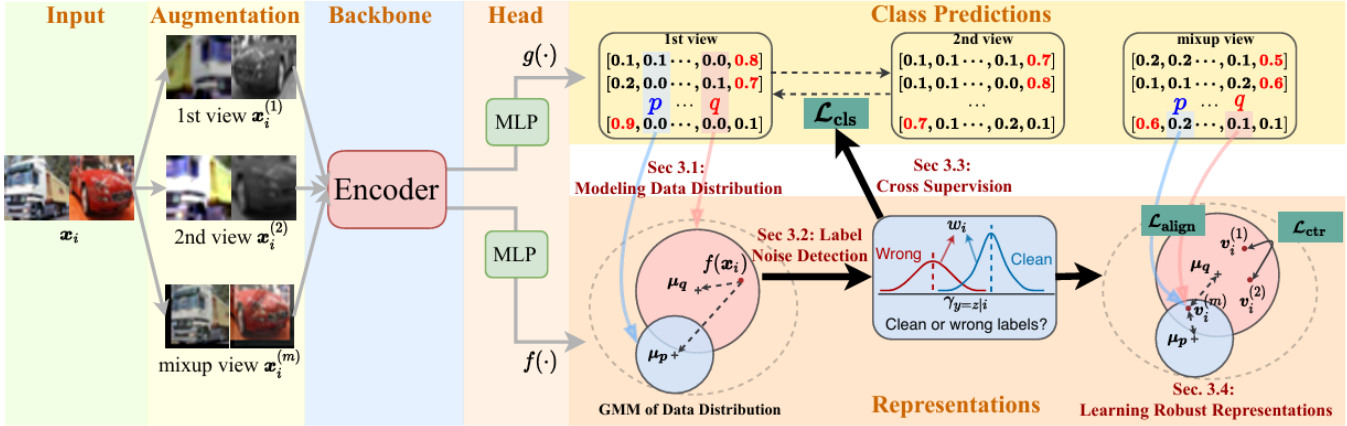

Fig. 1. Illustration of the proposed TCL. The networks g and f with shared encoder and independent two-layer MLP output the class predictions and representations. Then, TCL models the data distribution via a GMM, and detects the examples with wrong labels as out-of-distribution examples. To optimize TCL, these results lead to cross-supervision and robust representation learning.

where norm$(\cdot)$ is ℓ_2-normalization such that $\|\mu_k\|_2 = 1$.

B. Out-Of-Distribution Label Noise Detection

After building the connection between the latent variables z and labels y, we are able to detect the sample with wrong labels through the posterior probability in Eq. (2). We implement it as a normalized version to take into account the intra-cluster distance, which allows for detecting the samples with likely wrong labels:

$$\gamma_{ik} = \frac{\exp\left(-\frac{(v_i-\mu_k)^T(v_i-\mu_k)}{2\sigma_k}\right)}{\sum_k \exp\left(-\frac{(v_i-\mu_k)^T(v_i-\mu_k)}{2\sigma_k}\right)}. \tag{5}$$

Since ℓ_2-normalization has been applied to both embeddings v and the cluster centers μ_k, yielding $(v-\mu_k)^T(v-\mu_k) = 2 - 2v^T\mu_k$. Therefore, we can re-write Eq. (5) as:

$$\gamma_{ik} = p(z_i = k|x_i) = \frac{\exp(v_i^T\mu_k/\sigma_k)}{\sum_k \exp(v_i^T\mu_k/\sigma_k)}. \tag{6}$$

we define conditional probability to measure the probability of one sample with clean label:

$$\gamma_{y=z|i} = p(y_i = z_i|x_i) = \frac{\exp(v_i^T\mu_{z_i}/\sigma_{z_i})}{\sum_k \exp(v_i^T\mu_k/\sigma_k)}. \tag{7}$$

so the wrong label automatically is denoted by $(1-\gamma_{y=z|i})$. Two-components GMM for clean and wrong label is defined as follows:

$$p(\gamma_{y=z|i}) = \sum_{c=0}^{1} p(\gamma_{y=z|i}, c) = \sum_{c=0}^{1} p(c)p(\gamma_{y=z|i}|c), \tag{8}$$

C. Cross-supervision with Entropy Regularization

After the label noise detection, TCL leverages a similar idea to bootstrap the targets through the convex combination of its noisy labels and the predictions from the model itself:

$$\begin{cases} t_i^{(1)} = w_i y_i + (1-w_i)g(x_i^{(1)}) \\ t_i^{(2)} = w_i y_i + (1-w_i)g(x_i^{(2)}) \end{cases} \tag{9}$$

where $g(x_i^{(1)})$ and $g(x_i^{(2)})$ are the predictions of two augmentations, y_i the noisy one-hot label, and $w_i \in [0,1]$ represents the posterior probability as $p(c = 1|\gamma_{y=z|i})$ from the two-component GMM defined in Eq. (8). Guided by the corrected labels $\tilde{t}_i$, we swap two augmentations to compute the classification loss twice, leading to the bootstrap cross supervision, formulated as:

$$L_{\text{cross}} = \ell\left(g(x_i^{(1)}), \tilde{t}_i^{(2)}\right) + \ell\left(g(x_i^{(2)}), \tilde{t}_i^{(1)}\right). \tag{10}$$

where ℓ is the cross-entropy loss. This loss makes the predictions of the model from two data augmentations close to corrected labels from each other. In a sense, if $w_i = 0$, the model is encouraged for consistent class predictions between different data augmentations, otherwise $w_i = 1$ it is supervised by the clean labels. In addition, we leverage an additional entropy regularization loss on the predictions within a mini-batch $\mathcal{B}$:

$$L_{\text{reg}} = -\mathcal{H}\left(\frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}} g(x)\right) + \frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}} \mathcal{H}(g(x)) \tag{11}$$

where $\mathcal{H}(\cdot)$ is the entropy of predictions. The first term can avoid the predictions collapsing into a single class by maximizing the entropy of average predictions. The second term is the minimum entropy regularization to encourage the model to have high confidence for predictions, which was previously studied in semi-supervised learning literature.

D. Learning Robust Representations

To model the data distribution that is robust to noisy labels, we leverage contrastive learning to learn the representations of images. Specifically, contrastive learning performs instance-wise discrimination using the InfoNCE loss to enforce the model outputting similar embeddings for the images with semantic preserving perturbations. Formally, the contrastive loss is defined as follows:

$$L_{\text{ctr}} = -\log \frac{\exp\left(\frac{f(x^{(1)})^T f(x^{(2)})}{\tau}\right)}{\sum_{x\in S} \exp\left(\frac{f(x^{(1)})^T f(x)}{\tau}\right)} \tag{12}$$

where τ is the temperature and S is the B except $x^{(1)}$. $x^{(1)}$ and $x^{(2)}$ are two augmentations of x.

E. Total Loss

The overall training objective is to minimize the sum of all losses:

$$L = L_{\text{cls}} + L_{\text{ctr}} + L_{\text{align}}. \tag{13}$$

V. Learnings

This large-scale deep-learning project significantly enhanced my technical expertise and problem-solving skills. I gained a deeper understanding of how complex deep learning architectures function, their strengths, and their limitations. Working with real-world datasets taught me the importance of data cleaning, transformation, and analysis to ensure optimal model performance. Limited resources forced me to optimize training strategies and creatively tailor model architectures. Overcoming numerous challenges significantly improved my debugging and troubleshooting skills. Furthermore, I became highly adaptable, quickly learning new tools, libraries, and techniques as the project demanded.

VI. Results

Due to resource constraint, I could only train 91 epochs. This image shows the test accuracy, training accuracy and knn accuracy for 91 epochs:

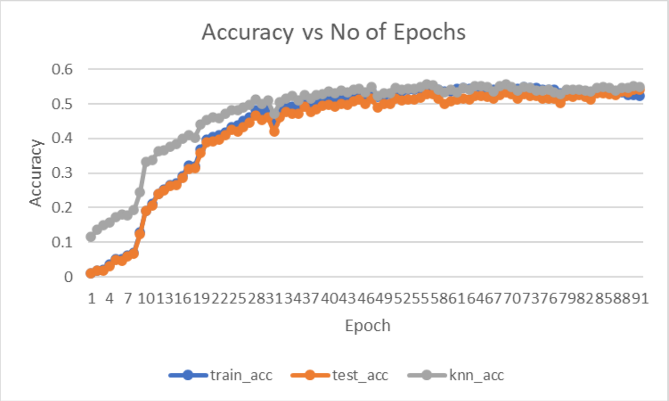

Fig. 2. Illustration of the proposed TCL. The networks g and f with shared encoder and independent two-layer MLP output the class predictions and representations. Then, TCL models the data distribution via a GMM, and detects the examples with wrong labels as out-of-distribution examples. To optimize TCL, these results lead to cross-supervision and robust representation learning.

VII. Conclusion

In this paper, we implemented a Twin Contrastive Learning model for learning from noisy labels. TCL can effectively detect the label noise and accurately estimate the true labels. In particular, TCL works effectively under extreme noise ratio(like 90%). In the future, it can be improved with semantic information for low noise ratios as well. Due to resource constraints, we could do experiment with the cifar100 dataset for 91 epochs only but this paper achieves 7.5% performance improvement under extremely 90% noise ratio.

References

[1] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, 2017.

[2] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Car los Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In Advances in Neural Infor mation Processing Systems, 2020.

[3] Xinshao Wang, Yang Hua, Elyor Kodirov, and Neil M Robert son. IMAE for noise-robust learning: Mean absolute error does not treat examples equally and gradient magnitude's variance matters. arXiv preprint arXiv:1903.12141, 2019.

[4] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 322 330, 2019.

[5] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L DMI: Anovel information-theoretic loss function for train ing deep nets robust to label noise. In Advances in Neural Information Processing Systems, 2019.

[6] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In Advances in Neural Information Processing Systems, 2018.

[7] https://www.cvat.ai/