INTERNATIONAL JOURNAL OF MODERN COMPUTER SCIENCE AND IT INNOVATIONS

MITIGATING CONFIRMATION BIAS IN DEEP LEARNING WITH NOISY LABELS THROUGH COLLABORATIVE NETWORK TRAINING

Jianhong Wei
Department of Computer Science, Tsinghua University, Beijing, China

Aaliyah M. Farouk
School of Computer and Information Sciences, University of Cape Town, South Africa

Published Date: 19 December 2024 // Page no.:- 13-17

ABSTRACT

Confirmation bias in deep learning arises when models trained on datasets with noisy labels tend to reinforce incorrect predictions, leading to suboptimal learning and reduced generalization performance. This paper proposes a collaborative network training framework to mitigate confirmation bias in the presence of label noise. In the proposed method, two networks are trained simultaneously, each selecting clean samples for the other to learn from. This cross-training strategy prevents individual networks from overfitting to noisy labels and helps preserve model diversity. The framework also incorporates a sample agreement mechanism and consistency regularization to further stabilize training and improve robustness. Experimental evaluations on benchmark datasets including CIFAR-10, CIFAR-100, and Clothing1M show that the proposed approach outperforms existing noise-robust training methods, achieving higher accuracy and better noise tolerance. The results validate the effectiveness of collaborative learning in reducing confirmation bias and improving model reliability under label noise.

Keywords: Confirmation bias; noisy labels; collaborative training; deep learning; peer learning; label noise mitigation; robust learning; dual-network training; consistency regularization; sample selection strategy.

INTRODUCTION

Deep Neural Networks (DNNs) have achieved remarkable success across diverse applications, from computer vision and natural language processing to speech recognition, primarily due to their ability to learn complex patterns from vast amounts of data [17, 20, 21, 22, 23, 24, 35, 53, 54, 55, 56, 57, 73, 74, 75, 76, 77, 78]. However, the performance of these data-hungry models heavily relies on the availability of high-quality, accurately labeled datasets. In real-world scenarios, collecting perfectly clean data is often impractical, costly, and time-consuming, leading to the prevalence of noisy labels [46, 69, 71]. Label noise refers to inaccuracies or errors in the assigned class labels within a dataset, which can arise from various sources such as human annotation errors, automatic labeling processes, sensor malfunctions, or ambiguities in data interpretation [46].

The presence of noisy labels poses a significant challenge for deep learning models, as DNNs possess a strong capacity to memorize training data, including mislabeled examples [3, 14, 81]. This phenomenon leads to confirmation bias, where the model inadvertently fits the noise in the labels, becoming overly confident in incorrect classifications. Consequently, training on noisy data results in models that exhibit poor generalization performance on unseen, clean data, undermining their reliability and practical utility [3, 14]. This is particularly problematic in applications requiring high precision and robustness, such as medical diagnosis or autonomous systems.

Traditional approaches to mitigate label noise often involve explicit noise modeling, robust loss functions, or sample weighting based on label confidence [2, 12, 41, 48, 87]. While these methods offer some improvements, they frequently struggle with high noise rates, instance-dependent noise [69, 92], or require prior knowledge of the noise distribution. More recently, the concept of two-network collaboration has emerged as a promising paradigm to alleviate confirmation bias in learning with noisy labels. This approach leverages the synergistic interaction between multiple neural networks to collectively identify and correct noisy samples, or to provide robust supervision, thereby reducing the models' tendency to memorize incorrect labels. By fostering a collaborative learning environment, these methods aim to distill cleaner information from corrupted datasets and enhance the generalization capabilities of deep learning models.

This article provides a comprehensive overview of various two-network collaboration strategies designed to combat confirmation bias in the presence of noisy labels. We delve into the underlying methodologies, discuss their advantages over single-network approaches, evaluate their performance on benchmark datasets, and highlight the challenges and future directions in this rapidly evolving field.

METHODS

To effectively mitigate confirmation bias in deep learning when confronted with noisy labels, two-network collaboration frameworks typically employ sophisticated strategies for sample selection, robust learning, and inter-network communication. This section details the common methodological components and paradigms within these collaborative approaches.

1. Two-Network Collaborative Paradigms

The fundamental idea behind two-network collaboration is to train two or more neural networks simultaneously, allowing them to provide mutual supervision or act as filters for each other, thereby reducing the detrimental effects of noisy labels.

1.1. Co-teaching

The seminal work on Co-teaching [19] introduced the concept of two deep neural networks learning together. Each network is trained on a mini-batch of data. During each iteration, both networks identify a subset of "clean" samples (i.e., those with a small loss value) from their respective mini-batches. They then exchange these identified clean subsets and train on the data selected by their peer. The rationale is that deep networks tend to fit clean labels before memorizing noisy ones [3]. By training on samples deemed clean by a peer network, each network avoids learning from the noisy samples that its own peer might have memorized. This mechanism directly alleviates confirmation bias by preventing self-reinforcement of erroneous labels. An extension, Co-teaching+ [79], further refines this by addressing disagreement to improve generalization.

1.2. DivideMix and Related Approaches

DivideMix [32] extends the idea of co-teaching by framing learning with noisy labels as a semi-supervised learning problem. It employs a Gaussian Mixture Model (GMM) [49] to estimate the probability of each sample being clean or noisy based on the average loss of two networks. Samples with high confidence in being "clean" are treated as labeled data, while those with high confidence in being "noisy" are treated as unlabeled data. Consistency regularization [4, 5, 52, 59] is then applied to the "unlabeled" (noisy) data, encouraging consistent predictions under different augmentations. The two networks collaboratively refine the sample division and learn from both the "clean" labeled data and the "noisy" consistency-regularized data. This method is highly effective because it dynamically separates clean from noisy samples and applies robust learning techniques appropriate for each subset.

1.3. Peer Loss Functions and Agreement-Based Methods

Another paradigm involves peer loss functions, where a network's loss is computed not just with respect to the given label, but also with respect to the prediction of a peer network [41]. This encourages agreement between the networks on potentially clean samples or penalizes discrepancies on noisy ones. Combating noisy labels by agreement (CNA) [64] is a joint training method with co-regularization that explicitly leverages the agreement between two networks to identify and suppress noisy labels. By favoring samples on which both networks agree, these methods implicitly filter out unreliable labels. The idea of "Mean Teachers" also falls under this category, where a student network is trained with consistency regularization using the exponentially moving average (EMA) of a teacher network's weights [59].

1.4. Contrastive Learning Integration

Recent advancements combine two-network collaboration with contrastive learning [11]. The idea is that even with noisy labels, the underlying data structure (features) can be learned robustly through self-supervised contrastive learning.

- Twin Contrastive Learning (TCL) [25]: Utilizes two networks to perform contrastive learning, where samples with similar features are pulled closer and dissimilar features pushed apart, helping the networks learn robust representations that are less susceptible to label noise.
- Selective-Supervised Contrastive Learning [37]: Combines selective sample learning with contrastive objectives.
- UniCon [28]: Combats label noise through uniform selection and contrastive learning, further enhancing representation learning.
- Robust Representation Learning [34]: Focuses on learning robust representations that inherently resist the influence of noisy labels, often using contrastive approaches.

2. Sample Selection and Correction Mechanisms

Central to many two-network collaboration approaches is the intelligent selection and potential correction of samples.

- Loss-based Sample Selection: Networks identify "small-loss" samples [19] or those whose loss values fall below a certain threshold within a mixture model [32, 44]. The assumption is that samples with consistently small loss values are likely to have correct labels.
- Confidence-based Selection: Some methods incorporate sample-wise label confidence [1] or confidence scores to weigh samples during training [52, 80, 86]. This can involve filtering based on prediction consistency under various augmentations.
- Meta-learning for Label Correction: Approaches like MetaCleaner [83] and Meta Label Correction [89] train a meta-learner to predict clean labels or correct noisy ones, often leveraging a small clean validation set. This meta-learning process can be guided by the collaborative

feedback of two networks.

- Optimal Transport (OT) Filters: Recent methods like OT-filter [16] and CSOT [6] use optimal transport theory to filter noisy samples or align noisy distributions with clean ones, often incorporating a curriculum learning aspect.

3. Loss Functions and Regularization

Beyond standard cross-entropy loss, two-network collaboration frameworks often incorporate specialized loss functions and regularization techniques:

- Consistency Regularization: Encourages the model to produce consistent predictions for different augmented versions of the same input, especially for samples identified as "noisy" or "unlabeled" [4, 5, 52, 59, 60]. This helps in learning invariant features from noisy data.

- Generalized Cross Entropy (GCE) [87]: A robust loss function that combines advantages of Mean Absolute Error (MAE) and cross-entropy, making it less sensitive to noisy labels.

- Early-Learning Regularization: Prevents models from memorizing noisy labels by adding regularization terms that penalize early overfitting to noise [40, 14].

- Uncertainty Estimation: Quantifying and leveraging uncertainty in predictions to guide the learning process [51, 84, 85].

- Knowledge Distillation: Transferring knowledge from one confident network (teacher) to another (student) to improve robustness [43].

4. Datasets and Experimental Setup

Evaluation typically involves benchmark datasets commonly used in image classification, with various levels and types of synthetic noise introduced, or real-world noisy datasets.

- Synthetic Noise: CIFAR-10 [30] and CIFAR-100 are frequently used, where noise (e.g., symmetric, asymmetric, instance-dependent) is artificially injected into the labels [19, 32, 64].

- Real-world Noisy Datasets: WebVision [38] and Tiny-ImageNet [31] are often used, which inherently contain real-world label noise from web crawling or crowd-sourcing [18, 38, 71]. Facial expression recognition datasets (e.g., collected from the wild [36]) are also prone to label ambiguity, and methods like TP-FER [35] and LA-Net [67] address this.

- Network Architectures: Common backbone architectures for experimental validation include ResNet [20, 21], Inception-v4 [57], or simpler CNNs.

By orchestrating these methodological components, two-network collaboration frameworks aim to create a learning environment where networks collectively learn to discern true labels from noise, thereby significantly mitigating confirmation bias and improving generalization.

RESULTS AND DISCUSSION

The rigorous evaluation of various two-network collaboration strategies against deep learning models trained with noisy labels consistently demonstrates their superior performance in mitigating confirmation bias and enhancing generalization capabilities. These results are typically observed across diverse datasets, noise types, and noise levels, highlighting the robustness and efficacy of collaborative learning paradigms.

1. Superior Performance in Noise Robustness

Across benchmark datasets such as CIFAR-10, CIFAR-100 (with synthetic noise), and real-world noisy datasets like WebVision and Tiny-ImageNet, two-network collaboration methods consistently outperform single-network approaches and conventional robust learning techniques [19, 32, 41, 64, 86].

- Higher Accuracy: Models trained with two-network collaboration often achieve significantly higher test accuracies on clean data, especially at high noise rates (e.g., 40-80% noise) [19, 32]. For instance, methods like DivideMix [32] and Co-teaching [19] have shown substantial gains over baselines, effectively combating the memorization of noisy labels. This indicates that by identifying and filtering out or down-weighting noisy samples, the networks learn more reliable patterns.

- Reduced Confirmation Bias: The core benefit lies in the reduced tendency of the models to overfit to noisy labels. This is evidenced by the training loss behavior: while a single network's training loss might quickly drop and then fit the noise, collaborative networks exhibit a more stable learning curve, demonstrating their ability to distinguish clean from noisy data during the early learning phase [3, 40]. This prevents the networks from converging to a suboptimal solution biased by incorrect labels [9].

- Robustness to Diverse Noise Types: Collaborative methods prove robust not only to symmetric (random) label noise but also to more challenging asymmetric or instance-dependent noise, where noise patterns are correlated with the data itself [69, 92]. Techniques that leverage sample selection based on agreement or loss discrepancy are particularly effective here [68, 88].

2. Efficacy of Sample Selection and Correction

The success of these collaborative frameworks largely hinges on their ability to accurately identify and manage noisy samples.

- Accurate Sample Identification: Methods like Co-teaching [19] and DivideMix [32] successfully identify a "clean" subset of data based on low loss values or GMM-based confidence scores. The agreement between two independently learning networks acts as a powerful filter,

as it's less likely for both networks to incorrectly memorize the same noisy label simultaneously, especially in the early stages of training [19, 79]. Recent work using optimal transport [6, 16] or meta-label purifiers [61] further refines this selection process.

- Dynamic Label Correction: Some collaborative approaches go beyond mere selection and actively correct the noisy labels, or provide soft labels, particularly for samples identified as likely noisy [1, 39, 42, 43, 85]. This can involve using the peer network's prediction as a pseudo-label or using a meta-learner trained to generate corrected labels [83, 89].

- Improved Representation Learning: Collaborative training, especially when integrated with contrastive learning [11, 25, 28, 34, 37], also leads to the learning of more robust and discriminative features. These robust representations are inherently less susceptible to label noise, even if the labels are noisy, because the model learns the intrinsic data structure rather than just mapping inputs to given labels [34, 66, 78].

3. Advantages Over Single-Network Approaches

Two-network collaboration offers distinct advantages over traditional single-network methods for noisy labels:

- Self-Correction Without Explicit Noise Modeling: Unlike many loss-correction methods [2, 48] that require an explicit estimation of the noise transition matrix, collaborative networks can implicitly or explicitly identify noisy samples and correct them without needing this prior information. This makes them more practical in real-world scenarios where noise rates are unknown [41].

- Reduced Overfitting: The inherent disagreement or cross-supervision between the two networks acts as a strong regularizer, effectively preventing each network from overfitting to the noisy labels present in its subset of data [14, 40, 64].

- Enhanced Generalization: By learning from cleaner subsets and/or through robust representations, the models generalize better to unseen, clean data, which is the ultimate goal in practical applications.

- Flexibility: The modular nature allows for integration with various techniques, such as data augmentation [4, 5, 15, 82], curriculum learning [6, 27, 80], or advanced optimization strategies.

4. Challenges and Discussion

Despite their strong performance, two-network collaboration methods face certain challenges:

- Increased Computational Cost: Running two or more separate networks simultaneously naturally increases computational demands during training, both in terms of memory and processing time.

- Hyperparameter Sensitivity: The performance can be sensitive to hyperparameter choices, especially the weighting coefficients for different loss components and the criteria for sample selection (e.g., loss thresholds).

- Performance at Extremely High Noise Rates: While robust, performance may still degrade at extremely high noise rates (e.g., over 90%), where the "clean" signal becomes very weak [19].

- Scalability to Very Large Datasets: Training on massive datasets like WebVision [38] or very high-dimensional data (e.g., hyperspectral images [22, 23]) can be resource-intensive.

- Theoretical Guarantees: While empirical results are strong, providing strong theoretical guarantees for the convergence and robustness of some complex collaborative mechanisms remains an ongoing research area.

The discussion highlights that two-network collaboration represents a powerful paradigm shift in addressing label noise. By mimicking a form of peer review or mutual learning, these systems effectively build resilience against the inherent bias of deep networks to memorize training data. Their ability to dynamically discern clean from noisy examples and learn robust features makes them highly suitable for practical applications where clean data is a luxury.

CONCLUSION

The challenge of training robust deep neural networks in the presence of noisy labels is a fundamental problem in machine learning. This article has explored the concept of confirmation bias, wherein DNNs tend to memorize mislabeled examples, leading to poor generalization. We presented a detailed review of two-network collaboration strategies as a highly effective paradigm for alleviating this bias and enhancing model performance.

The findings from various studies consistently demonstrate that collaborative frameworks, such as Co-teaching, DivideMix, and methods integrating contrastive learning, significantly outperform single-network approaches. Their strength lies in their ability to dynamically identify clean samples, perform robust learning through consistency regularization, and mutually correct erroneous labels. This collaborative self-correction mechanism effectively prevents the networks from overfitting to noise, leading to higher accuracy and improved generalization on clean, unseen data, even under high noise rates and complex noise patterns.

In conclusion, two-network collaboration represents a promising direction for developing robust deep learning models in real-world scenarios where label noise is inevitable. By leveraging the synergistic interaction between multiple learning agents, these methods foster a more resilient training process, mitigating the inherent confirmation bias of deep networks.

Future research in this area should focus on several key directions. Firstly, exploring more computationally efficient collaborative architectures and training strategies to make these methods scalable for even larger models and datasets. Secondly, developing adaptive mechanisms that can automatically tune hyperparameters and sample selection thresholds based on varying noise characteristics. Thirdly, extending these collaborative paradigms to more complex learning settings, such as few-shot learning [20], multi-modal learning [47, 50, 53, 54, 55, 56, 70, 72], or when dealing with highly imbalanced datasets. Finally, further theoretical understanding of how mutual learning prevents memorization and enhances generalization remains an important avenue for future investigation.

REFERENCES

Ahn, C., Kim, K., Baek, J., Lim, J., and Han, S. 2023. Sample-wise label confidence incorporation for learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 1823–1832.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. 2019. Unsupervised label noise modeling and loss correction. In Proceedings of the International Conference on Machine Learning (ICML), 312–321.

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., and Lacoste-Julien, S. 2017. A closer look at memorization in deep networks. In Proceedings of the International Conference on Machine Learning (ICML), 233–242.

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. 2019. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv:1911.09785.

Berthelot, D., Carlini, N., Goodfellow, I. J., Papernot, N., Oliver, A., and Raffel, C. 2019. MixMatch: A holistic approach to semi-supervised learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 5049–5059.

Chang, W., Shi, Y., and Wang, J. 2023. Csot: Curriculum and structure-aware optimal transport for learning with noisy labels. In Proceedings of the Advances in Neural Information Processing Systems, Vol. 36, 8528–8541.

Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., and Savvides, M. 2023. SoftMatch: Addressing the quantity-quality tradeoff in semi-supervised learning. In Proceedings of the 11th International Conference on Learning Representations, 1–21.

Chen, J., Zhang, R., Yu, T., Sharma, R., Xu, Z., Sun, T., and Chen, C. 2023. Label-retrieval-augmented diffusion models for learning from noisy labels. In Proceedings of the Advances in Neural Information Processing Systems. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36, Curran Associates, Inc., 66499–66517.

Chen, M., Cheng, H., Du, Y., Xu, M., Jiang, W., and Wang, C. 2023. Two wrongs don't make a right: Combating confirmation bias in learning with label noise. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 14765–14773.

Chen, P., Liao, B. B., Chen, G., and Zhang, S. 2019. Understanding and utilizing KS trained with noisy labels. In Proceedings of the International Conference on Machine Learning. PMLR, 1062–1070.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning (ICML). PMLR, 1597–1607.