**Appendix of IEEE TNNLS Submission "Federated Learning with Noisy Clients"**

# 1 Additional Experimental Results

## 1.1 Evaluation with modern model architectures

To demonstrate the efficacy of Fed-NCL on modern model architectures, we conducted a comprehensive experiment on CIFAR-10 using Transformers architecture. The Transformer architecture has been increasingly favored in recent literature due to its adaptability to heterogeneous data distributions and remarkable resilience to distribution shifts [1]. However, prior research on federated learning with noisy labels mainly focuses on convolutional neural networks (CNN) and the performance of modern architectures such as the Transformer is largely unknown [2–4]. To address this gap, we employed a standard implementation of Vision Transformers [5, 6] for image tasks, including image classification [7, 8], and compared the performances of Fed-NCL and FedAvg using various Bernoulli noise scenarios on CIFAR-10. Our findings, as presented in Tab.1, demonstrate that the Transformer architecture exhibits more robustness to clients with noisy labels in FL as compared to the CNN architecture. This robustness can be attributed to the self-attention mechanism of the Transformer which can learn the global pattern of the task [9], leading to less bias towards local patterns than the CNN. Despite the Transformer's robustness, our experiments reveal that Fed-NCL still outperforms FedAvg at different noise levels, indicating that the noise-robust layer-wise aggregation in Fed-NCL can lead to better weight distribution during model aggregation. These results demonstrate the effectiveness of Fed-NCL in modern model architectures and highlight the importance of noise-robust aggregation methods in federated learning.

Table 1: Test accuracy on ViT(s) on FedAVg and Fed-NCL with under various noise settings. The data was distributed across 5 clients in an independent and identically distributed (IID) manner under the Bernoulli noise scenario. The reported results include the best precision for each method.

| | IID with clean clients | | Bernoulli with $p = 0.6$ | | Bernoulli with $p = 0.7$ | |
|---|---|---|---|---|---|---|
| Methods | Central | FedAvg | FedAvg | Fed-NCL | FedAvg | Fed-NCL |
| ViT(S) | 97.17% | 98.50% | 96.95% | 97.30% | 97.08% | 97.65% |

## 1.2 Evaluation with Real World Human Annotated Noise

To show our method's effectiveness on real world noise scenarios, we conducted an experiment on CIFAR-10N [10] with Gaussian noise scenario. The CIFAR-10N is a variant of the popular CIFAR-10 dataset, consisting of 60,000 32x32 color images divided into 10 different classes for image classification. The training set of CIFAR-10N is augmented with human-annotated real-world noisy labels collected from Amazon Mechanical Turk. The level of noise in CIFAR-10N varies for each image, depending on its complexity and ambiguity, reflecting the inherent biases of human beings. This makes CIFAR-10N a more challenging and practical dataset for learning with noisy labels, compared to artificially generated label noise. In our experiment, we distributed the data to the clients in a non-iid setting with $\alpha_{\text{DIR}} = 10, p = 0.7$. This allowed us to validate

the effectiveness of our proposed approach when dealing with complex and diverse noisy data. The results, as shown in Tab.2, demonstrate that Fed-NCL outperforms other methods by up to 40% and 13% on average. The feature-dependent noise of CIFAR-10N presents a more challenging situation for the noise detector, making it difficult for some small-loss methods to detect the noisy clients. However, Fed-NCL shows consistent performance between human and synthetic label noise, as the noisy detection in Fed-NCL catches up with the divergence layer by layer, making it more robust and accurate than loss-based methods. Overall, our experiment on CIFAR-10N provides strong evidence of the effectiveness of Fed-NCL in real-world scenarios involving complex and diverse label noise.

Table 2: Accuracy of CIFAR-10N with different methods. The average accuracy of the last 10 rounds is reported.

| Method | FedAvg | FedProx | Trimm | FOCUS | Ours |
|---|---|---|---|---|---|
| Accuracy | 67.19% | 67.52% | 67.98% | 31.53% | 72.05% |

## 2 Proposed Federated Algorithms

To tackle the above challenges, we introduce Federated Noisy Client Learning (Fed-NCL), which effectively distinguishes noisy clients and intelligently mitigates their impact during the overall FL process. Algorithm 1 illustrates the full Fed-NCL algorithm. An overview of Fed-NCL, mainly contains the following three stages: 1) noisy client detection, 2) robust layer-wise adaptation aggregation, and 3) label correction. To identify noisy clients, the server calculates the reliability scores of the clients to statistically determine the noisy clients. After detecting the noisy clients, the server performs robust layer-wise adaptation aggregation, which jointly considers the model's layer divergence and the impact of noisy clients, to obtain a global model for the next round of local training. Finally, we correct the labels from the noisy clients to reduce their negative impact and extract more valuable features from them.

**Algorithm 1** Fed-NCL

**Input:** Local Epoch $E$, batch Size $B$, detection std $\beta$, confident threshold $\eta$ , label correction time factor $\alpha$
**Server:**
  **for** each communication round $t = 1, 2...T$ **do**
    $S_t \leftarrow$ (random select K out of N clients)
    **for** each client $c \in S_t$ **do**
      $\Theta_t^c, h_t^c \leftarrow$ **Client($\Theta_t^G$)**
    **end for**
    $Q_t \leftarrow$ **Calculate reliability score** using Eq.**??**
    $S_n, S_c \leftarrow$ **Detect noisy clients** using Eq.**??**
    Update label correction candidates $S_{corr}$
    $\Theta_{t+1}^G \leftarrow$ **Robust Layer-wise-Aggregation** using Eq.**??**
  **end for**
**Client($\Theta_t^G$):**
  **if** $c \in S_{corr}$ **then**
    $\widetilde{D}_c \leftarrow$ Label Correction using Eq.**??**
  **end if**
  **for** each local epoch $e$ from 1 to $E$ **do**
    $\mathcal{B} \leftarrow$ Randomly split local data $(\mathcal{X}, \mathcal{Y})$ into batches of size $B$
    **for** minibatch $(x_b, y_b) \in \mathcal{B}$ **do**
      $\Theta_t^c \leftarrow \Theta_t^c - \eta \nabla l(f(x_b; \Theta_t^c), y_b)$
    **end for**
    $h_t^c = \sum_{(x_i, y_i) \in D^c}^{N} \mathrm{CE}(f(x_i; \Theta_t^c), y_i)$
  **end for**
  **return** $\Theta_t^c, h_t^c$

# References


[1] L. Qu, Y. Zhou, P. P. Liang, Y. Xia, F. Wang, E. Adeli, L. Fei-Fei, and D. Rubin, "Rethinking architecture design for tackling data heterogeneity in federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 061–10 071.

[2] J. Xu, Z. Chen, T. Q. Quek, and K. F. E. Chong, "Fedcorr: Multi-stage federated learning for label noise correction," *arXiv preprint arXiv:2204.04677*, 2022.

[3] Y. Chen, X. Yang, X. Qin, H. Yu, B. Chen, and Z. Shen, "Focus: Dealing with label quality disparity in federated learning," *arXiv preprint arXiv:2001.11359*, 2020.

[4] T. Tuor, S. Wang, B. J. Ko, C. Liu, and K. K. Leung, "Overcoming noisy and irrelevant data in federated learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5020–5027.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[7] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[8] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[9] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *Advances in neural information processing systems*, vol. 32, 2019.

[10] J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, and Y. Liu, "Learning with noisy labels revisited: A study using real-world human annotations," *arXiv preprint arXiv:2110.12088*, 2021.