Experimenting with Information Dissimilarity for Knowledge Distillation at the Edge

Redacted for double-blind review

***Abstract*—Knowledge distillation (KD) is a key technique for transferring knowledge from a large, complex "teacher" model to a smaller, more efficient "student" model. KD is extensively used to facilitate knowledge transfer between Edge devices in distributed infrastructures. While Cross Entropy (CE) and Kullback-Leibler (KL) are commonly used in KD, this work investigates the applicability of loss functions based on underexplored information dissimilarity measures, such as Triangular Divergence (TD), Structural Entropic Distance (SED), and Jensen-Shannon Divergence (JS), for both independent and identically distributed (iid) and non-iid data distributions. The primary contributions of this study include an empirical evaluation of these dissimilarity measures within a decentralized learning context, i.e., where independent clients collaborate without a central server coordinating the learning process.**

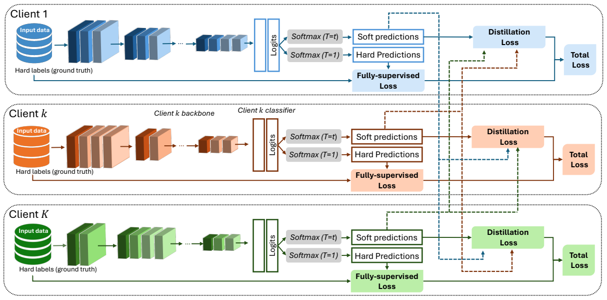


Fig. 1: KD-based decentralized network consisting of K clients, where distillation is performed using soft predictions for effective knowledge transfer.

I. Introduction

The integration of Artificial Intelligence in edge processing has led to the emergence of an interdisciplinary field known as *Distributed Intelligence* or *Edge Intelligence*, which aims to develop systems composed of software agents, robots, sensors, and computer systems that can collaborate effectively [1], [2], [3]. In this field, *Knowledge distillation* (KD) has been employed to facilitate knowledge transfer between edge devices, enhancing the development of more efficient and accurate models [4]. KD is a machine learning technique designed to transfer knowledge from a large, complex model (the *teacher*) to a smaller, more efficient one (the *student*) [5], [6], [7]. In addition to its primary role in model compression, it has started to find applications in other areas, including distributed intelligence [8] and continual learning [9].

In the KD-based distributed learning framework, *clients* exchange information to enhance their learning process, where each client operates as a learner and a source of knowledge for other clients. These clients are part of a decentralized system where no single model acts as the central teacher. Instead, each client trains on its local dataset and shares knowledge with others. As illustrated in Fig. 1, this information exchange is achieved through combining two types of losses. The first loss component, indicated as "fully-supervised loss", is usually the cross-entropy (CE) with "hard" targets derived by the ground-truth labels of the input samples. The second component is the "distillation loss" designed to ensure that each learning client mimics the output of other remote clients [10]. This loss is typically implemented by comparing the probability distributions of the models involved, where one model acts as the student and others take turns serving as teachers. This encourages the student's output probabilities to closely match those of the teacher. The model's output probabilities are typically computed using a softmax layer. Adjusting the softmax temperature during training has proven to be crucial in metric learning and distillation processes. In the context of distributed intelligence, this technique is also employed to generate soft predictions for effective distillation. Hence, the distillation loss is expressed as minimizing the gap between the soft predictions of one client with respect to the soft predictions of all other clients [11], [12], [13].

Given that the softmax function transforms an array of logits into an array of positive values summing to 1, various information dissimilarity measures can theoretically be used to implement the distillation loss. However, in practice, it is predominantly realized using CE, in addition to Kullback-Leibler (KL) Divergence, and Mean Squared Error (MSE) [14]. These methods have been extensively studied and proven effective for knowledge transfer in diverse machine learning tasks, while a wide range of information distance functions remain unexplored in the literature related to distributed learning.

In this work, we break new ground by investigating alternative dissimilarity measures – specifically, Triangular Divergence (TD), Structural Entropic Distance (SED), and Jensen-Shannon (JS) divergence – in the context of KD for decentralized learning scenarios. Recently, the correlations among these measures and the commonly used CE have been examined in [15] for independent and identically distributed (iid) data. Our work aims to expand the understanding of how these dissimilarity measures can enhance KD techniques, particularly in settings where data distribution may vary across learning clients (with a non-iid data distribution).

Our main contributions include designing a distributed KD environment suitable for investigating the aforementioned information dissimilarity measures and examining the performance of a set of clients by comparing pairwise distillation averaging among clients to the conventional peer-to-peer pairwise distillation, considering the various information dissimilarity measures.

II. Experimental Setup

Our analysis was conducted on a decentralized network consisting of three interconnected clients. We studied the effectiveness of different information dissimilarity measures (namely, CE, KL, SED, TD, JS) on distributed learning systems with different levels of data heterogeneity, ranging from scenarios where the data distribution is uniform across all clients (iid) to more extreme situations where each client focuses on its own specific tasks (non-iid). For this purpose, we used the CIFAR-10 [16] dataset and the SUN397 [17] dataset. We split the datasets into three subsets, one for each client. For the CIFAR-10, the iid distribution is obtained by shuffling and evenly splitting the entire dataset, ensuring each client has different samples. For the non-iid distribution across the clients, we followed the configuration in [4]. Each client C^k receives a subset $\{\ell_i\}$ of the labels, which are designated as primary labels for C^k. Labels not included in $\{\ell_i\}$ are considered secondary for C^k. Samples for each label ℓ are distributed randomly among clients, with a higher probability (1 + γ times greater) of being assigned to clients that have ℓ as a primary label. The parameter γ, referred to as dataset skewness, determines this distribution. In the experiments, we used $\gamma = 15$ for CIFAR-10 and $\gamma = 10$ for SUN397.

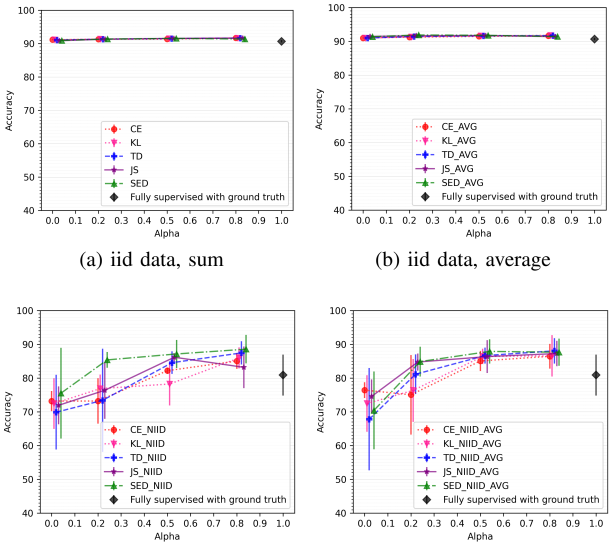
For defining the distillation loss $\mathcal{L}_{k,KD}$ we considered two alternatives:

- **Case 1**: The *sum* of pairwise dissimilarities between the current client's soft-prediction and remote client's soft-predictions.
- **Case 2**: A distillation loss based on the dissimilarity between the current client's soft-predictions and the *average* of soft-predictions from remote clients.

Performance evaluation was conducted using 10% of the entire data distribution for both iid and non-idd datasets. For each client, we computed the accuracy of the model. All models are based on ResNet18 [18] and are initialized with weights pre-trained on ImageNet, as provided by PyTorch. We also employ standard data augmentation techniques as recommended in the PyTorch documentation for ResNet18.

III. Results

Fig. 2a and Fig. 2b present the average accuracy on CIFAR-10, while varying the hyperparameter α, which control the amount of distillation loss ($\alpha = 0$ uses only distillation loss, $\alpha = 1$ uses only local loss). Our results indicate that KD does not significantly enhance overall accuracy when the input data is sufficient and balanced. Furthermore, all tested dissimilarity measures exhibited performance similar to CE. This observation is consistent across both cases when computing the distillation loss. Based on this observation, in iid settings, the choice of a dissimilarity measure may depend on implementation requirements, with a preference for computationally efficient measures such as TD. Fig. 2b shows that distillation with the average predictions of remote clients C^{Φ} results in similar accuracy to the sum of pairwise losses. This approach allows the computation of a single loss instead of multiple pairwise losses, potentially reducing computational complexity.



(a) iid data, sum

(b) iid data, average

(c) non-iid data, sum

(d) non-iid data, average

Fig. 2: CIFAR-10: Mean accuracy over three clients considering the *sum* of the distillation losses in the left-hand plots, and the *average* of remote predictions to compute the distillation loss in the right-hand plots.

In the case of non-iid distribution (Fig. 2c and Fig. 2d), the distillation process led to an increase in the average accuracy of the clients' models compared to the fully-supervised approach. This improvement is particularly noticeable for the value $\alpha = 0.5$. For this value, all measures show minimal variance among the three clients (as indicated by the vertical bars) except in 2c, where the KL provides a high variance compared to others. For $\alpha > 0$ values, minimal differences are observed between JS and SED when computing the distillation loss with the average of predictions generated by the remote clients, whereas CE and KL perform worse in case $\alpha = 0.2$. Furthermore, the average of the predictions obtained from remote clients, in Fig. 2d shows that for $\alpha = 0.2$, SED and JS already exhibit good performance. However, for $\alpha = 0.8$, all measures perform similarly, with KL having higher variance across clients. On the other hand, SED appears to be superior to other measures from $\alpha = 0.2$, providing minimal variance when considering the sum of distillation losses.

We also performed experiments in the non-iid scenario using the SUN397 dataset. In these experiments, adding more layers to the second head caused the model to overfit, showcasing an average accuracy of 48.33% compared to the first head, showcasing an average accuracy of 57.74% over all the clients. This confirms the argument made in [4] that when the client's training data is scarce, leading to model overfitting, communication between clients can enhance generalization and improve client's performance on their private tasks.

Regarding the performance of the different dissimilarity measures, CE and KL are outperformed by SED, TD, and JS distances for $\alpha = 0$ and $\alpha = 0.8$ when using the sum of distillation losses from each remote client. However, when using the average of remote predictions, the CE and KL perform worse for the values of $\alpha = 0$ and $\alpha = 0.2$. In other cases, all measures perform equally well.

References

[1] L. E. Parker, "Distributed intelligence: Overview of the field and its application in multi-robot systems." in *AAAI fall symposium: regarding the intelligence in distributed intelligent systems*, 2007, pp. 1–6.

[2] Y. Sahni, J. Cao, S. Zhang, and L. Yang, "Edge mesh: A new paradigm to enable distributed intelligence in internet of things," *IEEE access*, vol. 5, pp. 16 441–16 458, 2017.

[3] X. Liu, J. Yu, Y. Liu, Y. Gao, T. Mahmoodi, S. Lambotharan, and D. H.-K. Tsang, "Distributed intelligence in wireless networks," *IEEE Open Journal of the Communications Society*, vol. 4, pp. 1001–1039, 2023.

[4] A. Zhmoginov, M. Sandler, N. Miller, G. Kristiansen, and M. Vladymyrov, "Decentralized learning with multi-headed distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8053–8063.

[5] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 3 2015.

[6] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[7] J. Gou, X. Xiong, B. Yu, L. Du, Y. Zhan, and D. Tao, "Multi-target knowledge distillation via student self-reflection," *International Journal of Computer Vision*, vol. 131, no. 7, pp. 1857–1874, 2023.

[8] I. Bistritz, A. Mann, and N. Bambos, "Distributed distillation for on-device learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 593–22 604, 2020.

[9] A. Carta, A. Cossu, V. Lomonaco, D. Bacciu, and J. van de Weijer, "Projected latent distillation for data-agnostic consolidation in distributed continual learning," *Neurocomputing*, p. 127935, 2024.

[10] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1365–1374.

[11] A. Agarwala, J. Pennington, Y. Dauphin, and S. Schoenholz, "Temperature check: theory and practice for training models with softmax-cross-entropy losses," *arXiv preprint arXiv:2010.07344*, 2020.

[12] G. Aguilar, Y. Ling, Y. Zhang, B. Yao, X. Fan, and C. Guo, "Knowledge distillation from internal representations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 7350–7357.

[13] Z. Yang, A. Zeng, Z. Li, T. Zhang, C. Yuan, and Y. Li, "From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 185–17 194.

[14] T. Kim, J. Oh, N. Kim, S. Cho, and S.-Y. Yun, "Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation," *arXiv preprint arXiv:2105.08919*, 2021.

[15] R. Connor, A. Dearle, B. Claydon, and L. Vadicamo, "Correlations of cross-entropy loss in machine learning," *Entropy*, vol. 26, no. 6, 2024.

[16] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)," 2009.

[17] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," *Advances in neural information processing systems*, vol. 27, 2014.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.