Information Dissimilarity Measures in Decentralized Knowledge Distillation: A Comparative Analysis

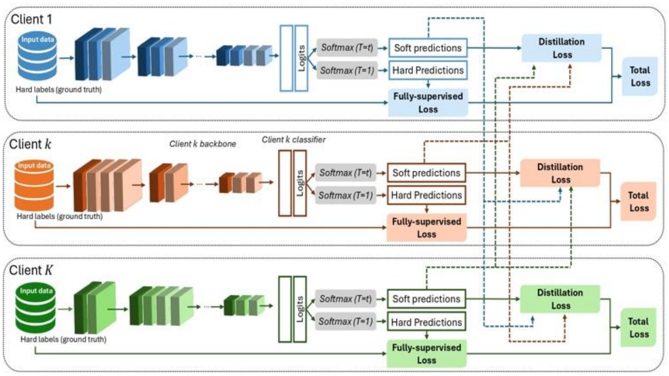
Joaquim Mbasa Molo[1,2], Lucia Vadicamo[1], Emanuele Carlini[1], Claudio Gennaro[1], Richard Connor[3]

[1]*Institute of Information Science and Technologies (ISTI), CNR, Pisa, Italy*
[2]*Department of Computer Science, University of Pisa, Pisa, Italy*
[3]*School of Computer Science, University of St Andrews, St Andrews, Scotland*

joaquim.molo@phd.unipi.it joaquimbasa/Distributed_KD_Information_Dissimilarity

UNIVERSITÀ DI PISA

University of St Andrews



KD-based Distributed Learning Framework

- Network of clients that cooperate for DNN training using **Knowledge Distillation** (KD)
- Each client acts as both learner (**student**) and source of knowledge (**teacher**) for others
- Each client C^k holds a local dataset D^k and a multi-head neural network $\mathcal{M}^k$, composed of :
  - **Backbone**: Extracts feature representations from input data
  - **Head 1**: Model $\mathcal{M}^k_{h1}$ (Backbone + Head 1) trained on local distribution D^k
  - **Head 2**: Model $\mathcal{M}^k_{h2}$ (Backbone + Head 2) trained on D^k using *knowledge distillation* from connected clients

Background: KD-based Information Exchange Mechanism (single teacher)

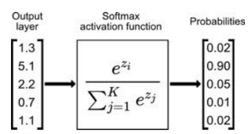
The student model is **trained** using two types of losses:

- **Fully-supervised loss** ($\mathcal{L}_{stu}$): Encourage the student's "hard" prediction to align closely with the ground-truth labels of the input samples
- **Distillation loss** ($\mathcal{L}_{KD}$): Encourages the student's output probabilities/representations to align closely with those of the teacher
  - Typically computed using Cross-Entropy (CE) and Kullback-Leibler (KL) Divergence (a wide range of information distance functions remains underexplored in distributed learning literature)

$$\mathcal{L} = \alpha\, \mathcal{L}_{stu} + (1-\alpha)\mathcal{L}_{KD}$$

$\mathcal{L}_{stu} = \boldsymbol{CE}(\boldsymbol{y}:\boldsymbol{p_S}(x, T=1)))$
y: "hard" targets from ground-truth
p_S: hard prediction of the student

$\mathcal{L}_{KD} = \boldsymbol{f}(\boldsymbol{p_T}(x, T=t):\boldsymbol{p_S}(x, T=t))$
p_T: soft prediction of the teacher
p_S : soft prediction of the student



$$\frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

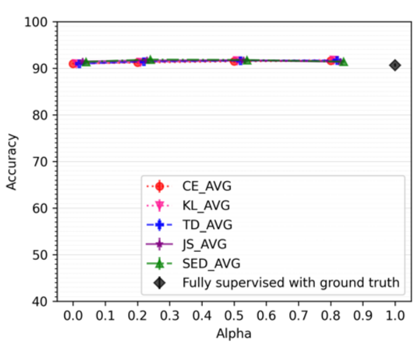
$$p = \frac{e^{z_i/T}(x)}{\sum_j e^{z_j/T}(x)}$$

$T \equiv temperature$

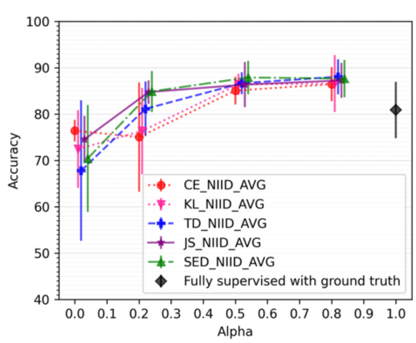
Our main contributions

- Since we have multiple teachers for a single student, we proposed two alternatives for computing the distillation loss.
  - **Sum** of pairwise dissimilarities between the current client's and remote clients' soft-predictions
  - Dissimilarity between the current client's soft-predictions and the **average** of remote clients' soft-predictions
- We performed experiments using three interconnected clients and tested different divergence functions for the KD loss:
  - Cross Entropy: $CE(\boldsymbol{q}:\boldsymbol{p}) = -\sum_{i=1}^{N} q_i \log p_i$
  - Kullback-Leibler Divergence: $KL(\boldsymbol{q}:\boldsymbol{p}) = \sum_i^N q_i \log\frac{q_i}{p_i}$
  - Jensen-Shannon Divergence: $JS(\boldsymbol{q},\boldsymbol{p}) = \frac{1}{2}\left(KL\left(\boldsymbol{q}:\frac{\boldsymbol{q+p}}{2}\right) + KL\left(\boldsymbol{p}:\frac{\boldsymbol{q+p}}{2}\right)\right)$
  - Structural Entropic Distance: $SED(\boldsymbol{q},\boldsymbol{p}) = \frac{C\left(\frac{q+p}{2}\right)}{\sqrt{C(\boldsymbol{p})C(\boldsymbol{q})}}$ $\quad C(\boldsymbol{p}) = b^{-\sum_{i=1}^{N} p_i \log_b p_i}$
  - Triangular Divergence: $TD(\boldsymbol{q},\boldsymbol{p}) = 1 - \sum_{i=1}^{N} \frac{2q_i p_i}{q_i + p_i}$
- We examined scenarios where the data is uniform across the clients, as well as cases in which the distribution is non-uniform

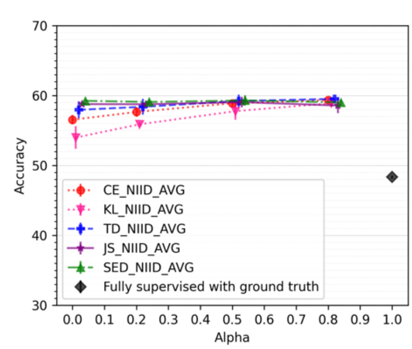
CIFAR-10 iid



CIFAR-10 non-iid



SUN397 non-iid



Conclusions

- We evaluated different information dissimilarity measures in a distributed KD setting across various data distributions
- The KD-loss based on the dissimilarity between the current client's soft-predictions and the *average* of soft-predictions from remote clients showed the best trade-off between accuracy and efficiency
- In the iid case, all measures have similar accuracy, however, the distance measures impact model training on *non-iid* data
- The commonly used Cross-entropy and Kullback-Leibler divergences are not always the most effective!