DarkDistill: Difficulty-Aligned Federated Early-Exit Network Training on Heteregeneous Devices

Lehao Qu[1], Shuyuan Li[2], Zimu Zhou[2], Boyi Liu[1,2], Yi Xu[1], Yongxin Tong[1]
[1]Beihang University, [2]City University of Hong Kong

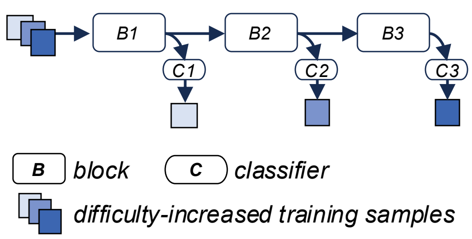
KDD 2025 Toronto

Abstract

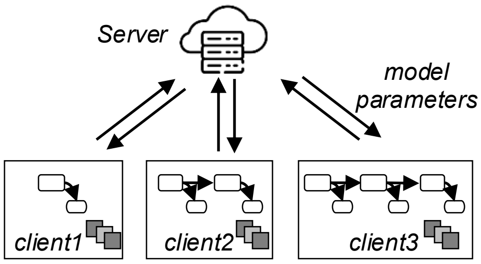
Early-exit networks (EENs), which adapt their computational depths based on input samples, are widely adopted to accelerate inference in edge computing applications. The effectiveness of EENs relies on difficulty-aware training, which tailors shallow exits for simple samples and deep exits for complex ones. However, existing difficulty-aware training schemes assume centralized environments with sufficient data, which become invalid with real-world edge devices.
In this paper, we explore difficulty-aware training in a federated manner, where EENs are collaboratively trained on heterogeneous devices. We observe the *cross-model exit unalignment phenomenon*, a unique problem when aggregating local EENs into a cohesive global model. To address this problem, we design a novel *Difficulty-Aligned Reverse Knowledge Distillation* scheme named DarkDistill that preserves the difficulty-specific specialization for aggregating heterogeneous local models. Instead of direct parameter averaging, it trains difficulty-conditional data generators, and selectively transfers generated knowledge of specific difficulty among matched exits of heterogeneous EENs. Evaluations show that DarkDistill outperforms the state-of-the-arts in various fine-tuning of EENs.

Introduction

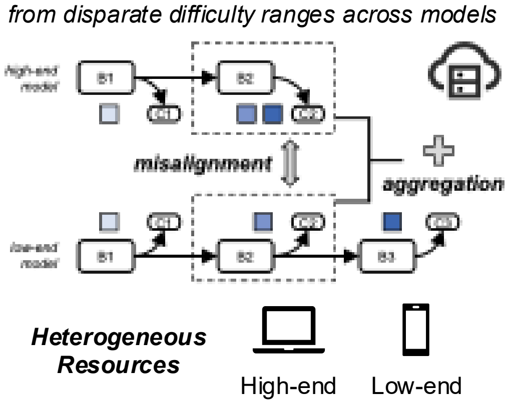
Early-Exit Network can adjust depth based on the difficulty of the input samples. *Easy (difficult)* samples terminate at *shallow (deep)* exits



Federated Learning EEN Training leverages the data knowledge of federated clients with heterogenous resources to train the *global EEN*



Cross-model Exit Unalignment
Exits at equivalent depths may handle samples from disparate difficulty ranges across models



New setting, new challenges!

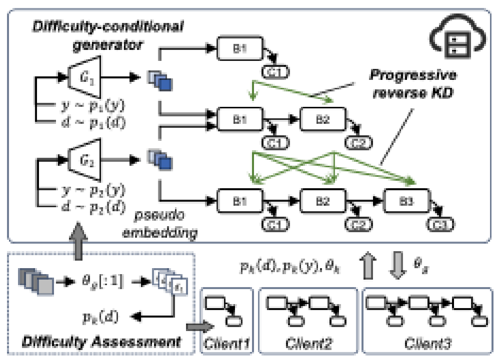
Methods

DarkDistill: Progressive Difficulty-Aligned Reverse Knowledge Distillation

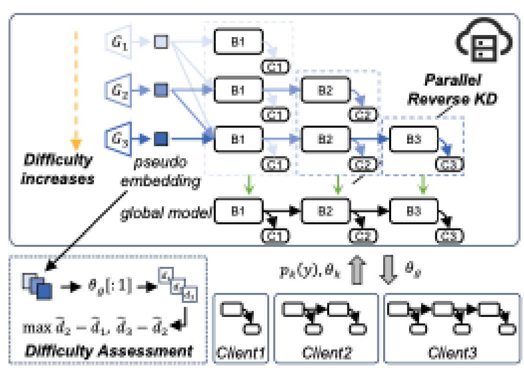
1. **Difficulty Assessment** evaluates the difficulty range of local data utilize its loss on global model

*2. **Difficulty-Conditional Generators** create pseudo data for specific difficulties, supporting the knowledge distillation process*

3. **Progressive Reverse KD** transfers knowledge from shallow to deep exits in adjacent layers across varied depth local EENs



DarkDistill-PL: Parallel Variant simultaneously distills the *ensemble knowledge* of all immediate knowledge to the global model parameterized by in an exit-wise manner



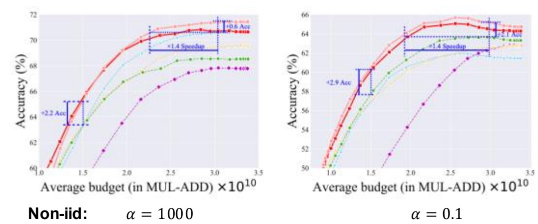
Difficulty-Aligned Knowledge Distillation

Experiments

Anytime Performance: Inference may terminate at any time. Show the average and variance of the accuracy across all exits on various datasets

Finetune	Difficulty-aware	Approach	CIFAR-100 [19] $\alpha = 0.1$	CIFAR-100 [19] $\alpha = 1$	CIFAR-100 [19] $\alpha = 1000$	SVHN [30]	SpeechCmds [44]
Full	None	ExclusiveFL	$26.60_{\pm 3.10}$	$49.96_{\pm 11.48}$	$41.58_{\pm 7.01}$	$85.28_{\pm 2.97}$	$87.00_{\pm 2.88}$
		InclusiveFL [26]	$40.10_{\pm 2.03}$	$58.83_{\pm 6.98}$	$61.40_{\pm 7.01}$	$82.95_{\pm 0.34}$	$91.90_{\pm 1.42}$
		ScaleFL [16]	$54.99_{\pm 10.61}$	$63.21_{\pm 9.14}$	$63.82_{\pm 9.87}$	$88.24_{\pm 0.78}$	$92.56_{\pm 0.26}$
		DepthFL [18]	$40.70_{\pm 1.57}$	$59.01_{\pm 5.18}$	$61.71_{\pm 5.73}$	$83.45_{\pm 0.43}$	$92.05_{\pm 0.60}$
		ReeFL [23]	$59.24_{\pm 8.00}$	$63.37_{\pm 7.72}$	$63.90_{\pm 8.68}$	$88.37_{\pm 1.27}$	$93.12_{\pm 1.14}$
	BoostNet [45]	ExclusiveFL	$48.68_{\pm 13.66}$	$57.57_{\pm 15.12}$	$58.65_{\pm 15.31}$	$87.30_{\pm 2.89}$	$91.07_{\pm 2.58}$
		InclusiveFL [26]	$57.10_{\pm 7.21}$	$62.96_{\pm 8.12}$	$64.01_{\pm 8.24}$	$87.86_{\pm 1.66}$	$92.91_{\pm 1.10}$
		ScaleFL [16]	$52.74_{\pm 13.82}$	$60.55_{\pm 11.93}$	$60.73_{\pm 10.80}$	$87.91_{\pm 0.77}$	$92.03_{\pm 0.37}$
		DepthFL [18]	$58.15_{\pm 6.73}$	$63.81_{\pm 6.34}$	$64.19_{\pm 6.73}$	$87.74_{\pm 1.01}$	$92.72_{\pm 0.64}$
		ReeFL [23]	$59.01_{\pm 7.98}$	$63.08_{\pm 9.03}$	$63.66_{\pm 7.31}$	$88.39_{\pm 1.28}$	$93.01_{\pm 1.18}$
		DarkDistill	$\underline{60.48}_{\pm 7.93}$	$\underline{64.50}_{\pm 7.97}$	**$65.67_{\pm 7.48}$**	$\underline{88.41}_{\pm 1.46}$	$\underline{93.31}_{\pm 1.13}$
		DarkDistill-PL	**$61.05_{\pm 8.19}$**	**$65.12_{\pm 7.02}$**	$\underline{65.49}_{\pm 7.88}$	**$88.48_{\pm 1.57}$**	**$93.42_{\pm 0.98}$**
LORA [13]	None	ExclusiveFL	$44.44_{\pm 18.61}$	$52.33_{\pm 18.56}$	$52.88_{\pm 18.17}$	$83.78_{\pm 4.43}$	$88.72_{\pm 3.15}$
		InclusiveFL [26]	$44.82_{\pm 23.36}$	$54.26_{\pm 21.38}$	$54.76_{\pm 21.37}$	$85.16_{\pm 5.31}$	$89.58_{\pm 3.04}$
		ScaleFL [16]	$22.17_{\pm 23.36}$	$30.85_{\pm 30.58}$	$32.58_{\pm 31.96}$	$76.42_{\pm 13.14}$	$58.82_{\pm 34.80}$
		DepthFL [18]	$52.17_{\pm 14.16}$	$57.09_{\pm 14.78}$	$57.63_{\pm 14.48}$	$85.69_{\pm 2.71}$	$90.11_{\pm 2.04}$
		ReeFL [23]	$52.32_{\pm 9.83}$	$57.74_{\pm 11.78}$	$58.16_{\pm 11.69}$	$85.54_{\pm 3.00}$	$89.56_{\pm 2.62}$
	BoostNet [45]	ExclusiveFL	$50.34_{\pm 13.63}$	$55.68_{\pm 15.33}$	$56.48_{\pm 15.33}$	$84.48_{\pm 3.88}$	$88.51_{\pm 2.26}$
		InclusiveFL [26]	$54.25_{\pm 11.78}$	$59.66_{\pm 11.72}$	$59.81_{\pm 11.66}$	$85.96_{\pm 2.50}$	$90.38_{\pm 2.10}$
		ScaleFL [16]	$40.46_{\pm 22.36}$	$47.18_{\pm 24.11}$	$48.26_{\pm 24.17}$	$81.70_{\pm 3.14}$	$80.19_{\pm 4.83}$
		DepthFL [18]	$55.85_{\pm 9.45}$	$60.95_{\pm 9.20}$	$61.45_{\pm 9.04}$	$79.90_{\pm 1.61}$	$90.93_{\pm 1.29}$
		ReeFL [23]	$51.57_{\pm 9.82}$	$58.04_{\pm 11.88}$	$58.62_{\pm 11.91}$	$85.44_{\pm 2.92}$	$89.40_{\pm 2.44}$
		DarkDistill	**$57.32_{\pm 11.91}$**	**$61.24_{\pm 11.06}$**	**$61.74_{\pm 11.42}$**	**$86.11_{\pm 2.16}$**	**$91.06_{\pm 2.08}$**

Budget Performance: Evaluate the accuracy of a batch across various budget. DarkDistill is faster and more accurate across multi-settings



Faster and more accurate

Theorem

Convergence Analysis DarkDistill converges in FL with heterogeneous clients

If the learning rate γ of local training satisfies $\frac{1}{T\sqrt{Q}} \le \gamma \le \frac{1}{6M^2LT}$, DarkDistill coverages to a neighborhood of a stationary point of **standard FL** as follows:

$$\frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}\|\nabla\mathcal{L}(\theta^q)\|^2 \le \frac{G_0}{\sqrt{Q}} + V_0 + \frac{H_0}{T} + \frac{I_0}{\sqrt{Q}}\sum_{q=1}^{Q}\mathbb{E}\|\theta^q\|^2$$

Explaination DarkDistill converges under a properly chosen learning rate γ, which can be practically set using the local epoch count T, total communication round Q, loss smoothness L, and largest exit number M

Guide to choose suitable lr