BoTTA: Benchmarking on-device Test Time Adaptation

Michal Danilowski[1], Soumyajit Chatterjee[2,3], Abhirup Ghosh[1,3]
[1]University of Birmingham [2]Nokia Bell Labs, Cambridge [3]University of Cambridge

I. Introduction

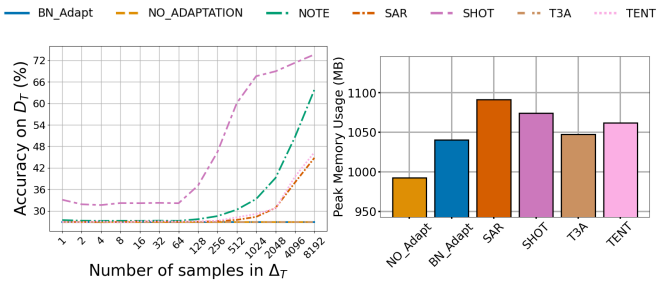
Deep learning models often rely on the assumption that test data follows the same distribution as training data. Let θ_S be a model trained on source data D_S. During deployment, the model is evaluated on target data D_T, which may follow a different and unknown distribution. This distribution shift can lead to significant performance degradation. Test-time adaptation (TTA) addresses this problem by updating model parameters during inference, using only unlabeled samples from D_T, without requiring access to the original training data D_S or any ground-truth labels. Due to its unsupervised nature, TTA is well-suited for on-device adaptation, where user privacy must be preserved and transmitting raw data to external servers is often undesirable. However, the constraints imposed by mobile and edge devices remain underexplored.

II. Problem Statement

In this presentation, we introduce BoTTA [2], our recent work on evaluating test-time adaptation (TTA) methods in realistic edge computing scenarios. TTA has gained attention as a way to improve model robustness under distribution shifts, but most existing evaluations focus on unconstrained settings and overlook practical deployment constraints. We identify four key challenges that affect TTA on an edge device: (1) A device will typically have access to a limited number of adaptation samples. This happens for the applications where the users actively participate in recording data, e.g. taking photos. as data is generated through user input or automated sensing, (2) limited number of categories in the adaptation data, as the user typically has limited exposure to the world, (3) D_T containing diverse distribution shifts, such as snow, rain or fog in outdoor photos depending on weather, and (4) overlapping shifts within a sample, for example, a foggy image also affected by motion blur. These factors are common in mobile and embedded deployments.

III. Evaluation Settings

We evaluate several representative SOTA TTA methods from each class of algorithms: softmax entropy minimization [1], sharpness aware entropy minimization [3], pseudo-labeling based [4], instance-aware batch normalization [5] and optimization-free [6]. We use two datasets (CIFAR-10C and PACS) and several model architectures (ResNet-26, ResNet-50, ViT). We test both in the server and run on real hardware platforms including Raspberry Pi 4B and Jetson Orin Nano. Alongside accuracy measured on different test datasets to evaluate the generalization of the adapted models, we report system-level metrics such as peak memory usage or computing power utilization.



(a) (b)

Fig. 1: (a) Test accuracy on D_T (CIFAR-10C Gaussian Noise domain, severity: 5) using ResNet-26 architecture. While accuracy increases with the increasing $|\Delta_T|$ (number of samples in on-device adaptation dataset ($\Delta_T \subseteq D_T$)), with data size below 64, none of the TTA methods works well. (b) peak memory consumption on Raspberry Pi-4B (8GB).

IV. Experimental Results

Our results show that many existing methods struggle when exposed to edge-specific constraints. For example, the most accurate method in Figure 1a - SHOT achieves $1.74\times$ accuracy gain when $|\Delta_T| = 256$ whereas it achieves $2.74\times$ gain when using a larger adaptation dataset ($|\Delta_T| = 8192$). Through the experiment on the real-life testbed, we observed that most of the TTA algorithms consume a significant amount of memory. Notably, in the experiments on Raspberry Pi in Figure 1b the most accurate TTA method across many settings, SHOT, consumes higher ($1.08\times$) peak memory compared to the base strategy of 'no adaptation.'. We also explore the idea of periodic adaptation instead of adapting on every inference step, which shows promise in balancing performance with practical feasibility. Through presentation, we aim to highlight open challenges and support the development of adaptation methods better suited for edge deployment.

References

[1] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization" in *International Conference on Learning Representations (ICLR)*, 2021.
[2] M. Danilowski, S. Chatterjee, A. Ghosh, "BoTTA: Benchmarking on-device Test Time Adaptation" in *arXiv preprint arXiv:2504.10149*, 2025.
[3] S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan, "Towards stable test-time adaptation in dynamic wild world" in *International Conference on Learning Representations (ICLR)*, 2023.
[4] J. Liang, D. Hu, and J. Feng, "Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation". in emphInternational Conference on Machine Learning (ICML) 2020 .
[5] T. Gong, J. Jeong, T. Kim, Y. Kim, J. Shin, and S.-J. Lee, "NOTE: Robust continual test-time adaptation against temporal correlation," *International Conference on Learning Representations (ICLR)*, 2022.
[6] Y. Iwasawa and Y. Matsuo, "Test-time classifier adjustment module for model-agnostic domain generalization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.