Efficient Test-Time Adaptation: Related Works

Juan Camacho Mohedano, Andrea De Carlo, Samuele Bolotta
DISI, University of Trento

***Abstract*—The ability of machine learning models to handle distributional shifts during inference is a critical challenge in real-world applications. Test-Time Adaptation (TTA) addresses this challenge by enabling models to dynamically adjust to unseen data distributions without requiring retraining or labeled target data. This review provides a comprehensive description of recent advances in TTA, categorizing methods into five primary strategies: model adaptation, normalization adaptation, sample adaptation, prompt adaptation, and inference adaptation. Each approach tackles unique aspects of distributional shifts, ranging from refining model parameters to leveraging external contextual information. Key methods, such as TENT for entropy minimization, TIP-X for modality alignment, and Training Dynamic Adapter (TDA) for real-time cache-based adaptation, are discussed alongside their strengths and limitations. Additionally, the impact of data stream characteristics, including IID and non-IID scenarios, is analyzed to highlight the importance of context-aware adaptation strategies. This review underscores the potential of TTA to enhance model robustness and efficiency across a range of domains, from autonomous driving to vision-language tasks, while identifying opportunities for future research.**

I. Introduction

The rapid evolution of machine learning models has brought forth a critical challenge: handling distribution shifts during test time. Test-Time Adaptation (TTA) represents an emerging paradigm that enables models to adapt dynamically to unseen data distributions without retraining or reliance on labeled target data. This field is particularly relevant for applications such as autonomous driving, medical imaging, and semantic segmentation, where unseen and unlabeled data are prevalent (Xiao et al., 2024).

TTA seeks to address the question: how can a model maintain robustness and efficiency in real-world conditions characterized by distributional shifts? To tackle this, researchers have proposed a range of methods categorized into five primary adaptation strategies: model adaptation, sample adaptation, prompt adaptation, normalization adaptation, and inference adaptation. Each of these categories addresses specific challenges and leverages distinct techniques to improve the model's ability to generalize.

II. Background

In test-time adaptation, we consider a source-trained model operating in a joint space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ represents the feature space and $\mathcal{Y}$ represents the label space. During training, we have access to source distribution samples $\mathcal{S} = \{p(\mathbf{x}_i^s, y_i^s)\}_{i=1}^N$, while at test time we encounter samples from a target distribution $\mathcal{T} = p(\mathbf{x}^t, y^t)$. The fundamental challenge arises from the distributional mismatch between source and target, expressed as $p(\mathbf{x}^t, y^t) \neq p(\mathbf{x}^s, y^s)$. A model f with parameters θ is trained on the source distribution (denoted as f_{θ^s}), and must adapt to make predictions on target samples without having seen any target data during training. The goal is to adapt f_{θ^s} to effectively handle target samples $\mathbf{x}^t$ at inference time, despite having no prior exposure to the target distribution during the training phase [10].

A. Distribution Shifts

Distributional shifts can be categorized into four fundamental types, each presenting unique challenges for adaptation:

- **Covariate shift:** Changes occur in the input distribution $p(\mathbf{x})$ while the conditional distribution $p(y|\mathbf{x})$ remains unchanged. Formally, $p(\mathbf{x}^t) \neq p(\mathbf{x}^s)$ but $p(y|\mathbf{x}^t) = p(y|\mathbf{x}^s)$. For example, models trained on images captured in sunny conditions must process images taken in snow.
- **Label shift:** The distribution of labels $p(y)$ changes while the conditional input distribution $p(\mathbf{x}|y)$ stays constant, expressed as $p(y^t) \neq p(y^s)$ but $p(\mathbf{x}|y^t) = p(\mathbf{x}|y^s)$. For instance, a visual recognition model trained on a balanced dataset of animal species encountering a region where certain species are far more common than others.
- **Concept shift:** The relationship between input and labels $p(y|\mathbf{x})$ changes due to noise or inconsistent labeling practices, represented as $p(y^t|\mathbf{x}^t) \neq p(y^s|\mathbf{x}^s)$.
- **Conditional shift:** Both $p(\mathbf{x})$ and $p(y|\mathbf{x})$ shift simultaneously, as in autonomous driving systems adapting between countries with different road rules and visual environments. This represents the most challenging scenario as it combines multiple types of shifts.

Most TTA methods, including those under model adaptation and normalization adaptation, focus primarily on addressing covariate shifts, as they represent the most common scenario in real-world applications. However, a growing body of work addresses more complex hybrid shifts that combine multiple types of distribution shifts [8]. Understanding these distinctions is crucial for developing robust adaptation strategies and selecting appropriate TTA methods for specific applications.

B. Data Stream Characteristics

In test-time adaptation, the characteristics of the incoming data stream critically influence the design and effectiveness of adaptation strategies. In the Independent and Identically Distributed (IID) setting, target samples $\mathbf{x}^t$ at time t and their corresponding ground-truth labels y^t (unknown to the

learner) are drawn independently from a time-invariant distribution $\mathcal{P}_T(x, y)$ [3]. This assumption simplifies the adaptation process because models can rely on the consistent statistical properties of the target distribution. Adaptation in IID settings is often computationally efficient, with minimal need for dynamic updates, as the distribution remains stable throughout inference. Such scenarios are more theoretical or occur in controlled environments where external variables are constrained.

In contrast, real-world data streams are predominantly non-IID, where samples follow a time-dependent distribution $\mathcal{P}_T(x, y|t)$, where t represents the temporal index [3]. Non-IID streams introduce complexities as consecutive samples are often temporally correlated, meaning that they are dependent on each other in ways that reflect the underlying context. For instance, in autonomous driving, object encounters will be dominated by cars while driving on highways but less so in downtown areas where pedestrians and bikes are more common. Similarly, in human activity recognition, some activities last for a short term (e.g., falling down), whereas certain activities last longer (e.g., sleeping). These conditions require adaptation mechanisms that go beyond static adjustments, employing strategies that track, anticipate and respond to temporal changes in the data. Recent works have proposed various approaches to handle non-IID streams, such as Instance-Aware Batch Normalization for detecting non-IID data through variance analysis [3] and Distribution Alignment for guiding test-time features back towards source distributions [9].

Furthermore, the level of drift—whether it is slow and predictable or rapid and unpredictable—affects the choice of adaptation strategy. Most existing TTA methods yield commendable results in ideal test data streams where batches are independently sampled, but they falter under more practical test data streams that are non-IID [9]. Slow drift might allow for batch-wise updates to model parameters, while rapid, unpredictable drift demands real-time adjustments to ensure performance does not degrade. The computational cost and latency associated with these strategies also vary, and balancing efficiency with adaptability remains a key challenge [11]. Understanding whether the incoming data stream adheres to IID assumptions or presents non-IID characteristics is therefore foundational to designing effective test-time adaptation pipelines tailored to specific applications.

C. Test-Time Adaptation

Test-time adaptation operates in two distinct phases. In the first phase, a model f_θ is trained on source distribution samples $\mathcal{S} = \{p(\mathbf{x}_i^s, y_i^s)\}_{i=1}^N$ in the joint space $\mathcal{X} \times \mathcal{Y}$. Once trained, this yields source-optimized parameters θ^s. In the second phase, the adaptation phase to the test time, the model encounters samples $\mathbf{x}^t$ from a target distribution $\mathcal{T} = p(\mathbf{x}^t, y^t)$ that differs from the source $(p(\mathbf{x}^t, y^t) \neq p(\mathbf{x}^s, y^s))$. The key innovation of TTA is that it adapts f_{θ^s} during inference to handle this distribution shift, modifying the model's behavior without requiring prior access to target domain data. This adaptation process occurs simultaneously with inference, allowing the model to adapt specific components to better handle samples from the previously unseen target distribution [10].

III. Approaches

Test-time Adaptation encompasses multiple approaches differentiated by the specific model components they adapt during inference. We categorize these approaches based on the type of updated component:

A. Model Adaptation

Model adaptation focuses on directly modifying the model's learnable parameters during test time. This approach involves fine-tuning or updating the model weights to align with the target distribution, allowing for a direct adjustment of the model's predictive capabilities.

Test-time entropy minimization (TENT) adapts models by fine-tuning batch normalization layers to minimize prediction uncertainty [8]. By reducing entropy, TENT sharpens predictions, making the model more confident in its outputs. For instance, an ambiguous image initially predicted as 50% cat, 30% dog, and 20% fox might, after adaptation, result in 90% cat, 5% dog, and 5% fox. This method is highly relevant to test-time adaptation (TTA) as it provides a simple yet effective way to improve performance without requiring labeled data. However, TENT struggles in highly uncertain scenarios, limiting its applicability in more ambiguous or complex shifts.

AdaContrast builds on the pseudo-labeling paradigm, refining predictions by leveraging contextual relationships in the feature space [1]. It employs contrastive learning to align similar features and separate dissimilar ones, ensuring that predictions are informed by neighboring data points. AdaContrast is particularly important in fine-grained classification tasks where relationships in the feature space provide critical information. Its ability to dynamically structure the feature space makes it a robust tool for TTA, although its reliance on nearest-neighbor search introduces computational complexity.

B. Normalization Adaptation

Normalization adaptation tries to overcome distribution shifts working on the statistics of the norm layers. In the literature this has been done with a linear combination of source and target statistics or inferring them with a meta learner.

NOTE [3] introduces an Instance-Aware Batch Normalization that manages to overcome the limitation of BN under distribution shift detecting non-IID data through variance analysis and correcting the normalization statistics for inference, and a Prediction-Based Reservoir Sampling for managing non-IID data streams and adapting gradually the statistics.

While effective, these methods are not model-agnostic, limiting their application to models with native normalization layers.

C. Sample Adaptation

An approach to avoid changing model parameters and requirements on the architecture is working on samples before feeding them into the model. Generative models are very common in this approach.

An example is [5], which achieves state-of-the-art in the domain generalization benchmarks by training a neural network to construct a label-preserving manifold so that a generative model maps samples to this manifold in a clustered manner, enabling effective classification.

Once trained, generative models are robust and stable, requiring no fine-tuning or additional data. However, their reliance on iterative updates per sample during inference reduces efficiency compared to other methods.

D. Prompt Adaptation

Another way to deal with large multimodal models is optimizing the prompts given in input at test time.

TPT (Test-time Prompt Tuning) [6], inspired by MEMO [12], pioneered this field, learning prompts via data augmentation and entropy marginalization. One drawback is that it may suffer from overconfidence in uncertain scenarios.

ZERO [2] acknowledge this problem pruning misleading confidence information from the augmented predictions with a softmax temperature set to the limit of zero.

RLCF [14] address the problem by combining reinforcement learning with a CLIP reward for more robust and reliable prompts.

Despite its promise, prompt adaptation faces challenges in generalizability, as prompts often overfit to specific tasks, and minimizing entropy over augmented views assumes that they are IID, but this is not always true if they are generated from the same stem image [2].

E. Inference Adaptation

Inference adaptation modifies the inference process to enhance real-time performance and reduce computational costs, making it particularly suited for applications requiring dynamic responses to test-time data.

Training Dynamic Adapter (TDA) introduces a flexible caching mechanism that adapts dynamically to incoming test data [4]. It employs two types of caches to handle varying levels of prediction confidence. The positive cache reinforces confident predictions, ensuring their reliability through repeated validation. The negative cache captures moderate-confidence samples, focusing on what the sample is not (e.g., "not a cat") to refine subsequent predictions. TDA's ability to efficiently manage certainty and ambiguity makes it highly relevant for test-time adaptation (TTA). It outperforms methods such as Test-Time Prompt Tuning (TPT) in both accuracy and computational efficiency, making it an excellent choice for real-world, resource-constrained scenarios.

TIP-Adapter employs a pre-built cache containing key-value pairs of image embeddings and their associated labels [13]. During testing, the model queries this cache to refine its predictions. This approach enhances efficiency by eliminating the need for backpropagation, relying instead on similarity-based retrieval for adaptation. While TIP-Adapter's static cache offers computational simplicity, it limits flexibility in dynamic environments, highlighting a trade-off between efficiency and adaptability. Despite this limitation, TIP-Adapter provides a lightweight and effective solution for scenarios where rapid inference is essential, complementing the dynamic capabilities of methods like TDA.

Another innovative approach is TIP-X, which addresses a unique challenge in vision-language models: the intra-modality gap, where image-only and text-only embeddings fail to align effectively [7]. TIP-X recalibrates intra-modal distances using Kullback-Leibler (KL) divergence, dynamically adjusting embedding relationships to reflect test-time data. This recalibration leverages external support sets, such as LAION-5B, to incorporate additional contextual information. By doing so, TIP-X enhances alignment across modalities, reducing the risk of misclassifications caused by embedding misalignment. This method is particularly critical for TTA as it bridges the gap between image and text modalities, ensuring more robust and accurate predictions. TIP-X's reliance on dynamic adjustment and external datasets makes it a powerful tool for improving adaptation efficiency and addressing modality gaps in real-world applications.

IV. Conclusions

Test-Time Adaptation (TTA) is an emerging paradigm that addresses the critical challenge of adapting machine learning models to unseen data distributions without requiring retraining or labeled target data. This paper reviewed the major approaches in TTA, categorized into model adaptation, normalization adaptation, sample adaptation, prompt adaptation, and inference adaptation. Each method offers unique advantages and faces specific limitations, highlighting the trade-offs between computational efficiency, robustness, and generalization.

Key contributions, such as TENT's entropy minimization, TIP-X's modality alignment, and TDA's dynamic caching mechanism, demonstrate the potential of TTA to improve model performance in real-world applications ranging from autonomous driving to vision-language tasks. Additionally, the importance of understanding data stream characteristics, particularly in non-IID scenarios, underscores the need for context-aware adaptation strategies.

While significant progress has been made, challenges remain in developing generalizable, scalable, and efficient TTA methods. Future research should focus on creating more model-agnostic solutions, improving adaptation in highly uncertain environments, and integrating domain knowledge to further enhance the adaptability of machine learning models. By addressing these challenges, TTA has the potential to become a cornerstone of robust, real-world AI systems.

References

[1] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation, April 2022. arXiv:2204.10377 [cs].

[2] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models, 2024.

[3] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: Robust Continual Test-time Adaptation Against Temporal Correlation, January 2023. arXiv:2208.05117.

[4] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models, March 2024. arXiv:2403.18293 [cs].

[5] Prashant Pandey, Mrigank Raman, Sumanth Varambally, and Prathosh Ap. Generalization on Unseen Domains via Inference-time Label-Preserving Target Projections. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12919–12928, Nashville, TN, USA, June 2021. IEEE.

[6] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models, September 2022. arXiv:2209.07511.

[7] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models, August 2023. arXiv:2211.16198 [cs].

[8] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization, March 2021. arXiv:2006.10726 [cs, stat].

[9] Ziqiang Wang, Zhixiang Chi, Yanan Wu, Li Gu, Zhi Liu, Konstantinos Plataniotis, and Yang Wang. Distribution alignment for fully test-time adaptation with dynamic online data streams, 2024.

[10] Z Xiao. Beyond model adaptation at test time: A survey.

[11] Zehao Xiao and Cees G. M. Snoek. Beyond Model Adaptation at Test Time: A Survey, November 2024. arXiv:2411.03687 version: 1.

[12] Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test Time Robustness via Adaptation and Augmentation, October 2022. arXiv:2110.09506.

[13] Renrui Zhang, Zhang Wei, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification, July 2022. arXiv:2207.09519 [cs].

[14] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with clip reward for zero-shot generalization in vision-language models, 2024.