

Persistent Test-time Adaptation in Recurring Testing Scenarios

Trung-Hieu Hoang ¹, Duc Minh Vo ², Minh N. Do ^{1,3}

¹ Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, USA

² The University of Tokyo, Japan

Best paper award – Community Track, 1st Workshop on Test-Time Adaptation: Model, Adapt Thyself! (MAT), Conference on Computer Vision and Pattern Recognition (CVPR) Workshop 2024.

NEURAL INFORMATION PROCESSING SYSTEMS

3 VinUni-Illinois Smart Health Center, VinUniversity

Context: Motivation for Test-time Adaptation (TTA)

Let's look at a real-world scenario: deploying a machine learning (ML) model for traffic surveillance camera system

Unforeseen circumstances can introduce *domain-shift* and *severely reduce* ML model's performance at test-time

How can we fix it?

- Making the ML model *learnable* at *test time*
- Utilizing *unlabeled* data at test time for adaptation
- TTA has been showing many "good" results!

Background: Test-time Adaptation for ML Model Deployment

Test-time Adaptation (TTA): TTA operates on an ML classifier $f_t: \mathcal{X} \rightarrow \mathcal{Y}$, parameterized by $\theta_t \in \Theta$ *gradually changing* over time.

Does the model performance/adaptability persist after a long time adapting to multiple environments?

What could go wrong?

Online testing stream: The model explores an online stream of testing data $X_t \sim \mathcal{P}_t$ for adapting itself $f_{t-1} \to f_t$ (self-supervised learning) before predicting $\hat{Y}_t = f_t(X_t)$.

Unfortunately, can not be guaranteed … we call it "TTA model collapsing"

Benchmark: Recurring Test-time Adaptation D_{e} outside en T_{e} at the environments Λ necurring rest-unse Ad

value of the category distribution (*p*⁰ term) which typically

Overview: Persistent Test-time Adaptation (PeTTA) $\bigcap_{x \in \mathcal{X}} \mathcal{L}_1$ of $\bigcap_{x \in \mathcal{X}} \mathcal{L}_2$ and $\bigcup_{x \in \mathcal{X}} \mathcal{L}_1$ 10Department of Electrical Engineering, University of India-Champaign and International A ²The University of Tokyo, Japan $x^2 + y^2 = 0$ entries than *Pt*(*y*), making the optimization step over *PM^t* more desirable. From Thm. 1, *M* moderates the extreme step *t*, evaluated on class *y* is defined as: *n* (rei IA) *Xt*, samples in this batch are selectively updated to a memototion *(*PoTTA) stances of *Xt*⁰ *, t*⁰ *< t* in the previous steps). By keeping a balanced number of samples from each category, distribu-

where *µ^y*

⁰ and ⌃*^y*

⁰ are the pre-computed empirical mean and

tion *P^M^t* (*y*) of samples in *M* is expected to have less zero

step *t*, evaluated on class *y* is defined as:

Theoretical Analysis: ε -Gaussian Mixture Model Classifier (ε -GMMC) LI (122 is exceeding to the unit due to the unit due to the unit of the unit due to the valii Analysis: E-Odussian iviixtt suggests a post-processing step for a post-processing step for adaptation (without up- $\begin{array}{ccc} \texttt{A} & \texttt{A}$ under this low intra-batch diversity (*|Yt|* ⌧ *|Y|*) situation can slowly degenerate the model (Boudiaf et al., 2022).

Goal: Simulating a simple yet representative failure case of TTA for theoretical analysis ting a simple vet representative failure case of TTA for theoretical ana achieving an adaptable continual test-time training. 4.1. Repeating TTA and Model Collapse

 16 Theoretical *ε*-Gaussian Mixture Model Classifier <mark>(*ε*-GMMC)</mark>

- Data stream: $\mathcal{X} \times \mathcal{Y} = \mathbb{R} \times \{0,1\}$ and the underlying joint distribution $P_t(x, y) =$ $p_{y,t} \mathcal{N}(x;\mu_y,\sigma_y^2)$ with $p_{y,t}=\Pr(Y_t=y)$ - true label and $\hat{p}_{y,t}=\Pr(\hat{Y}_t=y)$ - predicted label \mathcal{L}_{max} $\left(x;\mu_{\nu},\sigma_{\nu}^{2}\right)$ with $p_{\nu,t}=\Pr(Y_{t}=y)$ - true label and $\hat{p}_{\nu,t}=\Pr(\hat{Y}_{t}=y)$ - predicte $\frac{1}{2}$ $\frac{1}{2}$ maps an input image *x* 2 *X* to a category (label) *y* 2 *Y*. Let dation (or model collapse), we propose a *new testing sce*iying joint distribution $P_t(x, y) = 0$ et al. $p_{y,t} = \ln(t_t - y)$ - predicte
- Task: predicting X_t was sampled from cluster 0 or 1 (negative or positive) \overline{a} icting X_t was sampled from cluster 0 or 1 (negative or positive) ing TTA performs *revisiting the previous distributions K*
- ا را 134 *v* modifies the interpretation of \hat{Y} , and \hat{Y} . *Po* cultural production and a • Procedure: *pseudo-label* \hat{Y}_t *prediction and a mean-teacher update*

ous visits. For example, a sequence with *K* = 2 could be *P*¹ ! *P*² ! *···* ! *P^D* ! *P*¹ ! *P*² ! *···* ! *PD*. A formal definition of model collapse:

Repeating TTA. To study the gradual performance degra-

nary n a nodel collapse <i>narty and to be col- nodel is said to be collapsed from step $\tau \in \mathcal{T}, \tau < \infty$ *if there exists a non-empty subset of categories* $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$ *such that* $\Pr\{Y_t \in \tilde{\mathcal{Y}}\} > 0$ *but the marginal* $\Pr{\{\hat{Y}_t \in \tilde{\mathcal{Y}}\}}$ *converges to zero in probability:* lim $\lim_{t \to \tau} \Pr{\{\hat{Y}_t \in \tilde{\mathcal{Y}}\}} = 0.$ *lapsed from step* $\tau \in \mathcal{T}, \tau < \infty$ *if there exists a non-empty* \hat{u} is marginal $1 \cdot 1 \cdot t \in \mathcal{Y}$ f converges to zero in probability.

- "Noisy" pseudo-label predictor: The *predictor is perturbed* for retaining a false negative rate (FNR) of $\varepsilon_t = \Pr\{Y_t = 1 | \hat{Y}_t = 0\}$ to simulate undesirable effects of the TTA testing stream 143
143 **Y** 1 146 a qual a la la la maggiata mente a sua stat seudo-laber predictor. The *predict* rate (FNR) of $\varepsilon_t={\rm Pr}\{Y_t=1| \hat{Y}_t=0\,\}$ to simulate undesirable effects of the TTA testing stre: **t** (Vang et al., 2023; Dobler et al., 2023; Dobler et al., 2023; Dobler et al., 2023; Dobler et al. ≥023; Dobl Here, upon collapsing, a model tends to *ignore* all categories s perturbed for retaining a false neg $F = F \cdot F$ $MUSy$ pseudo-label piedictor. The pied predictions are zeros, the FNR also increases at every step r is perturbed for retaining a false negat
- ·hi 148 simple 11A model will be collaps first, the teacher model in the previous step introduces a $\overline{}$ Collapsing behavior varies across datasets and the adaptainclude a predictor for producing pseudo-label *Y*ˆ*^t* (Eq. 1), and a • How this simple TTA model will be collapsed? *GMMC model, with Assumption 1, if* lim *ing), the cluster 0 in GMMC converges in distribution to a*

retaining a false negative rate of ✏*^t* to simulate undesirable effects

r

Key Analysis Results 15 Pseudo-label Predictor *X^t* ✏*t* $-$ Kev \mathcal{L}_{G} (\mathcal{L}_{G} \mathcal{L}_{G} p_{max} and eventually reachest possible for p $\frac{1}{2}$ algorithms (e.g., category-balanced sampling strategory-balanced sampling strategory-balanced sampling strateg

167
167 - Paul VIII
167 - Paul VIII

*µ*ˆ*y,t* = (1 ↵)ˆ*µy,t*¹ + ↵E*^P^t ^Xt|Y*ˆ*^t* if *Y*ˆ*^t* = *y* \overline{e} -*GMMC model, with Assumption 1, if* $\lim \hat{p}_{1,t} = 0$ (*c* $\left($ **Lemma 2** (ϵ -GMMC After Collapsing). *For a binary k*^{I}_{*d*}^{*l*} \rightarrow ^{*l*}</sub>^{*l*} \rightarrow ^{*l*}^{\rightarrow}^{*l*} \rightarrow *l*</sub>^{*l*} \rightarrow *l*^{*l*} \rightarrow *l*</sub>*ld*_{*l*}*distribution do a single-cluster GMMC with parameters: to a single-cluster GMMC with parameters:* \sim severe distribution shifts or low intra-batch cat- ϕ *e-GMMC* model, with Assumption 1, if $\lim_{t\to\infty}\hat{p}_{1,t} = 0$ (col-

assumed to be static (defined below) and the pseudo-label **Corollary 1 (A Condition for** ϵ **–GMMC Collapse).** *With* fixed p_0 , α , μ_0 , μ_1 , ϵ – GMMC is collapsed if there exists a sequence of $\{\epsilon_t\}_{\tau-\Delta_{\tau}}^{\tau}$ ($\tau \geq \Delta_{\tau} > 0$) such that: ✏*^t* = Pr*{Y^t* = 1*|Y*ˆ*^t* = 0*}* (5) $d_{c-1}^{0\to 1}$ $\geq \epsilon_t > 1 - \frac{u_{t-1}}{1}, \quad t \in [\tau - \Delta_\tau, \tau].$ $p_1 \geq \epsilon_t > 1 - \frac{d_{t-1}^{0 \to 1}}{u_{t-1}}$ $\frac{t-1}{t}$

Assumption 1 (Static Data Stream). *The marginal distri-* μ_1 β ; μ_2 β α β *bution of the true label follows the same Bernoulli distribution* Ber(p_0)*:* $p_{0,t} = p_0$, $(p_{1,t} = p_1 = 1 - p_0)$, $\forall t \in \mathcal{T}$. $\begin{bmatrix} \text{non } \text{Der}(P_0) & P_0, & P_0, & P_1, & P_1 \end{bmatrix}$. The model collapse happens of P_1 . The model collapse \mathcal{P}_0 , \vert Assumption 1 (Static Data Stream). The marginal d dise \int *tion* Ber(p₀); $p_{0,t} = p_0$, ($p_{1,t} = p_1 = 1 - p_0$), $\forall t \in$

when this condition holds for a sufficient long period. The sufficient long period period. The sufficient long period.

h i

 $t\rightarrow \tau$ $\frac{1}{2}$ neters. $i\dot{r}$

securive steps, the simplicity of G allows solving the simplicity of G allows solving the simplicity of G

Lemma 1 (Increasing FNR). *Under Assumption 1, a binary* ϵ -GMMC would collapsed (Def. 1) with $\lim \hat{p}_{1,t} = 0$ $t\rightarrow \tau$ *(or* lim $t\rightarrow \tau$ $\hat{p}_{0,t} = 1$ *, equivalently) if and only if* $\lim_{h \to 0}$ $t \rightarrow \tau$ $p_{0,t} = 1$, equivalently) if and only if $\lim_{t \to \tau} \epsilon_t = p_1$. **1** (Increasing FNR). Under Assumption 1, a bi- \overline{M} $\frac{1}{2}$ **include a** predictor *for a product produce in a product* $\frac{1}{2}$ *, a product* $\frac{1}{2}$ ary ϵ -GMMC would collapsed (Def. 1) with $\lim_{t\to\tau} p_{1,t} = 0$ \overline{c} the testing stream in TTA, making \overline{c} and the collapse. $\begin{bmatrix} \n\frac{1}{2} & \frac{1}{2} & \frac{1$ $\lim_{t \to \tau} p_{0,t} = 1$, equivalently) if and only if $\lim_{t \to \tau} \epsilon_t = p_1$. Optim (for finding ✓⁰ **emma 1 (Increasing FNR).** Under Assumption 1, a bi- $\lim_{t \to \tau} \epsilon_t = p_1.$

Increasing the false-negative rate leads to model collapse $\frac{1}{2}$ 133 *Figure 2.* Diagram of our proposed ✏-perturbed binary Gaussian ne false-negative rate leads to model collapse
———————————————————— = Pr(*Y*ˆ*^t* = *y*) denote the marginal *^y}^y*2*^Y* is initialized at *t* = 0. For the conhe talse-negative rate leads to model colla_l:

178

h

*^L*CLS ⇣

 \mathcal{A}

⌘i

Theorem 1 (Convergence of ϵ **-GMMC).** For a bi **GMMC** model, with Assumption 1, let the distance from $\hat{\mu}_{0,t}$ $\left| \begin{array}{c} \end{array} \right|$ fixed toward μ_1 is $d_t^{0\rightarrow 1} = |\mathbb{E}_{P_t}[\hat{\mu}_{0,t}] - \mu_1|$, then: 181 tual TTA algorithms (e.g., category-balanced sampling Theorem 1 (Convergence of ϵ -GMMC). *For a binary* ϵ *toward* μ_1 *is* $d_t^{0\to1} = |\mathbb{E}_{P_t}[\hat{\mu}_{0,t}] - \mu_1|$ *, then:*

. Why collapsing? retaining a false negative rate of ✏*^t* to simulate undesirable effects ✓⁰ ⁼ *{µy,* ²

y,t). The goal is estimating a egory diversity of episodic/practical TTA *both result in an increase in the error rate of the predictor*. Instead of di-

 \overline{a} • After collapsing? verging to a single-cluster model (insteac $\overline{}$ Lemma 1 states the negative correlation between *p*ˆ1*,t* and and possible possible process. The highest possible \overline{P} Converging to a *single-cluster* model (instead of 2) *happens when µ*ˆ0*,t moves toward µ*1. We next investigate t_{M} and conditions for the conditions for the convergence. Theorem 1 (Convergence of ✏GMMC). *For a binary* ✏*-*

> $\sigma_0^2 + p_1 \sigma_1^2 + p_0 p_1 (\mu_0 - \mu_1)^2$. *^t* 's con-

 $\ell > 1 - \frac{1}{\lfloor t \rfloor}$, $t \in [T - \Delta_T, T].$ $t \in [\tau - \Delta_{\tau}, \tau].$

$\overline{2}$ d35L
—— *.* **t** are non-decreasing that limits are non-decreasing that limits $\frac{1}{2}$

$$
\mathcal{N}(\hat{\mu}_{0,t}, \hat{\sigma}_{0,t}^2) \stackrel{d}{\to} \mathcal{N}(p_0\mu_0 + p_1\mu_1, p_0\sigma_0^2 + p_1\sigma_1^2 + p_0p_1(\mu_0 - \mu_1)^2).
$$

study on ✏pertubed GMMC (✏GMMC), where *py,t* is *···* ✓*^t*¹ ✓*^t ···* `onditions and factors that contribute to t predictions are zeros, the FNR also increases at the FNR also increases at the FNR also increases at the FNR a
Eventy steps at every step in the FNR also increases at the FNR also increases at the FNR also increases at the on GMMC and TTA. GMMC first implies an *equal prior* μ distribution by continuate forms in α and α a pseudo after collapsing \overline{a} states the resulting \overline{a} Cluster 0 now *covers the whole data distribution* (and aslikelihood estimation *ft*(*x*) = argmax*^y*2*^Y* Pr(*x|y*; ✓*t*) with **happens when** *n***0**, *n***0**, *n***1.** We next investigate *n n***1.** We next investigate *n* Londitions and factors that contribute to the ϵ *µ*ˆ*y,t*¹ otherwise • How? Conditions and factors that contribute to the model collapse

The model collapse in Lemma 2) if *d*⁰!¹

out loss of generality, we study a particular *increasing type II collapse of* ✏*-GMMC*. By flipping the true positive pseudo Under the static data stream assumption $|\mu_0 - \mu_1|$ $|\mu_0 - \mu_1|$ Corollary 1 introduces a condition ✏-GMMC collapse.

$$
d_t^{0 \to 1} - d_{t-1}^{0 \to 1} \leq \alpha \cdot p_0 \cdot \left(|\mu_0 - \mu_1| - \frac{d_{t-1}^{0 \to 1}}{1 - \epsilon_t} \right).
$$

 $F₂ct₀$ *^N* (ˆ*µ*0*,t,* ˆ² ⁰*,t*) *d.* ! *N* (*p*0*µ*⁰ + *p*1*µ*1*,* Factors contributing to the model collapse: 188 steps, the simplicity of GMMC allows solving the Optim \sim From Thm. 1, we observe that the distance *d*⁰!¹ virting to the moder conapse. study on ✏pertubed GMMC (✏GMMC), where *py,t* is ntributing to the model collapse: $\hphantom{\ddots}$

- distribution of the true label *Y^t* and pseudo label *Y*ˆ*t*. $\mathcal{G}(\mathcal{G})$ tribution by construction t Lemma 2 states the result in the result of the result.
Lura difference hetween two categories the nature difference between two categories ($|\mu_0 - \mu_1|$); *r* actors contributing to the moder conapse.
(i) Data-dependent factors: the prior data distribution (p_0), $\frac{1}{101}$ <u>an</u> **a-dependent factors:** the prior data distribu **1**
1*PPnP<i>d***_{***F***}** nature difference between two categories (| μ_{0} -*(i)* χ ^t is similar to χ ^t *denendent factors* the prior data distributed *effects of the testing stream* on the predictor. The difference between two categories (
- $\binom{1}{i}$ gies in [15, 54]). Thus, it simplifies *f^t* into a maximum likelihood estimation *ft*(*x*) = argmax*^y*2*^Y* Pr(*x|y*; ✓*t*) with (iii) Algorithm-dependent factors: update rate (α) , the false *happens when it is a to ach step (* ε *_t)</sub>* $\qquad \qquad$ $t \to c$ negative rate at each step (ε_t) <u>The update of $\frac{1}{2}$ </u> *y y* ∗ $\overline{\mathsf{r}}$ *actors*: update rate (α), the f $\frac{1}{\sqrt{2}}$ $\frac{y}{x}$ = $\frac{y$ ontinin-dependent factors. update fate (α), t
ative rate at each step (s.) *fixed fixed p <i>fixed p <i>f*_c, *p d i*, *d i*, *sequence of {*✏*t}*⌧ be the false negative rate (FNR) of the model at step *t*. Withgories (*|µ*0*µ*1*|*); and (ii) *algorithm-dependent factors*: the update rate (↵), the FNR at each step (✏*t*). ✏-GMMC analy-

assigning label 0 for all samples). Furthermore, *collapsing*

⌧⌧ *(*⌧ ⌧ *>* ⁰*) such that:*

gies in [15, 54]). Thus, it simplifies *f^t* into a maximum

*µ*ˆ*y,t* =

We then obtained the following theoretical results: include a predictor for producing pseudo-label *Y*ˆ*^t* (Eq. 1), and a *GMMC model, with Assumption 1, if* lim r *p*ˆ1*,t* = 0 *(collapsing), the cluster 0 in GMMC converges in distribution to a* $\frac{1}{2}$ σ potained the following theoretical rest

assigning label 0 for all samples). Furthermore, *collapsing*

*^t < d*⁰!¹

*^t*¹ . The model collapse happens

t!⌧

a single cluster with parameters stated in Lemma 2. In the perturbed-free, GMMC converges to the true data distribution. (c) Distance

variations of PeTTA: without (w/o) regularizations of PeTTA: without (w/o) regularization term *R(*

PeTTA *(ours)* 23.7 23.1 22.8 22.6 23.0 22.6 22.8 22.7 23.2 23.1 23.2 23.1 22.9 23.1 22.8 22.8 22.7 22.9 23.5 23.6 23.0 Table 2. Average classification error of the task CIFAR-100 ! CIFAR-100-C in *episodic TTA* setting. of model should follow a similar predictions from a side on the predictions from a side of model should follow density function of the two clusters after convergence versus the true data distribution. The initial two clusters of ✏-GMMC collapsed into t ach column on these plots shows the histogram of model prediction (class labels are color-cod ' *Table 4.* Average (across 20 visits) classification error of multiple equal number of images for 10 classes. Hence, predictions from an ideal model should follow a uniform distribution. Table 1. Average classification error of the task CIFAR-10 ! CIFAR-10-C in *episodic TTA* setting. For all tables in the remaining of this

S imulation Results: Collapsing Behavior of ε -GMMC

9

Proposed Approach: Persisting Test-time Adaptation (PeTTA) p_{e} the divergence of p_{e} or p_{e} or p_{e} or p_{e} or p_{e} the output probability Pr(*y|x*; ✓*t*) and Pr(*y|x*; ✓0), we sug-Proposod Approach: Porcisting Tos step should be concentrated near the vectors of the initial terms of the initial terms of the initial terms of

- Now we introduce our *Persistent TTA (PeTTA)* approach. Further inspective Theorem inspective φ_{θ_t} is the deep-reature extraction. running mean of feature vector $\bm z\colon \{\hat\mu_t^{\nu}\}_{\nu\in\mathcal I}$ in which $\widehat\mu_t^{\nu}$ is exponential moving average updated with the value of vector **z** if $f_t(x) = y$ try in *z* and TTA approaches *learn to align feature vectors* f f_t , let $\mathbf{z} \,=\, \phi_{\theta_t}\left(\boldsymbol{x}\right)$. Keeping track of a collection of • Notation: With ϕ_{θ_t} is the deep-feature extractor of f_t , let $z = \phi_{\theta_t}(x)$. Keeping track of a collection of the **protation.** Write φ_{θ_t} is the deep reature extractor or θ funning mean or leature vector $\boldsymbol{z} \colon \{\mu_t\}_{y \in \mathcal{Y}}$ in which I reases in the p_{θ_t} and doop reason of the *y*^y and *u*^y and *n* $\sum_{y \in y}$ running mean of feature vector $\mathbf{z} \colon \{\hat{\mu}_t^{\mathcal{Y}}% (\theta)\}_{t=1}^{N}$ $\{\mathcal{C}^{\mathcal{Y}}_{t}\}_{y \in \mathcal{Y}}$ in which $\widehat{\bm{\mu}}^{\mathcal{Y}}_t$ is exponential moving average updated with the
- ed. Concess to the subsetstript of $f(x)$ is the value of ϕ_{θ} , from ϕ_{θ_0} , and adjust the adaptation objective corresponding *n*, and adjust the adaptation objective conesponentially ϕ Key Idea: Sensing the divergence of $\phi_{\theta_{t}}$ from $\phi_{\theta_{0}}$, *y* **Example 2.1.** The Constanting in Eq. 3. The constant of the from d_{θ} . • Key Idea: Sensing the divergence of ϕ_{θ_t} from ϕ_{θ_0} , and adjust the adaptation objective correspondingly
- $\mathcal{L}(\mathcal{A})$ is a proper adjustment toward toward the TTA algorithm t • With μ_0^{ϵ} , Σ_0^{ϵ} are pre-computed on the source dataset, we can: \mathbf{v}^{t} , \mathbf{v}^{t} and \mathbf{v}^{t} are proposition on the religion of the running of the section of the run in \mathbf{v}^{t} $\boldsymbol{\mu}_0$ $\boldsymbol{\mu}_0^{\mathsf{c}}$, $\boldsymbol{\Sigma}_0^{\mathsf{c}}$ are pre-computed on the source datase **Wit** $\mu_0^{\mathfrak{c}}$ $\mathbf{\Sigma}_0^t$ are pre-computed on the source • With μ_0^t , Σ_0^t are pre-computed on the source dataset, we can:

 $\frac{1}{4}$ Consider the divergence from θ

(2) Adaptive Learning Rate and Regularization

orated" version of the regular mean-teacher update model $\overline{D_{\alpha}TT\Lambda}$ is an "olaborato constant over the international chapter of the i **P L**AL(*X*^{*I*} α ^{*y*} **PeTTA** is an "elaborated" version of the regular mean-teacher update model ₉ ر
. which is equivalent to minimizing the KL divergence PeTTA is an "elaborated" version of the regular mean-teacher update model

(1) Sensing the divergence from
$$
\theta_0
$$

\n
$$
\gamma_t^y = 1 - \exp\left(-(\hat{\mu}_t^y - \mu_0^y)^T (\Sigma_0^y)^{-1} (\hat{\mu}_t^y - \mu_0^y)\right)
$$
\n
$$
\bar{\gamma}_t = \frac{1}{|\hat{\mathcal{Y}}_t|} \sum_{y \in \hat{\mathcal{Y}}_t} \gamma_t^y, \quad \hat{\mathcal{Y}}_t = \left\{\hat{Y}_t^{(i)} | i = 1, \dots, N_t\right\}
$$
\n
$$
\text{PerTA}
$$
\n
$$
\mathcal{A}_t = \bar{\gamma}_t \cdot \lambda_0, \qquad \alpha_t = (1 - \bar{\gamma}_t) \cdot \alpha_0,
$$
\n
$$
\theta_t = \text{Optim } \mathbb{E}_{P_t} \left[\mathcal{L}_{\text{CLS}} \left(\hat{Y}_t, X_t; \theta'\right) + \mathcal{L}_{\text{AL}} \left(X_t; \theta'\right)\right] + \lambda_t \mathcal{R}(\theta')
$$
\n(3) **Another Loss**
\n
$$
\mathcal{L}_{\text{AL}}(X_t; \theta) = -\sum_{y \in \mathcal{Y}} \Pr(y | X_t; \theta_0) \log \Pr(y | X_t; \theta)
$$
\n
$$
\text{(1 performance)}
$$
\n(1 collapse prevention)

Qualitative Results of PeTTA on CIFAR10-C

 $\overline{}$

4 ل

We qualitatively compare the performance of PeTTA (Persistent Test-time Adaptation) and RoTTA (Robust Testtime Adaptation [Yuan, 2023]) and analyze the model collapse on CIFAR10-C dataset

collapsed. Edges denote the average cosine similarity of feature vectors (source model), only the highest similar pairs are shown.

a few categories into a few categories into a few categories (e.g., and the categories and the model should follow
Any distribution of images for 10 classes. Hence-predictions from an ideal model should follow a uniform di 40 ו equal number of images for 10 classes. Hence, predictions from an ideal model should follow a uniform distribution.
Professional member of images for 10 classes. Hence, predictions from an ideal model should follow a unifo

Quantitative Results of PeTTA & Ablation Studies \bigcap ilantitative Results of PeTTA & Ahlatid Hill the seaso o collage categories into a few categories (e.g., *0: airplane, airplane, airplane, airplane, airplane, airplane, airplane, airplane,* denote the average cosine similarity of the average cost similar pairs (source model), only the shown. Best viewed in color. The highest similar pairs are shown. Best viewed in color. Best viewed in color. Best viewed in c

pas in recurring - $\overline{}$ \blacksquare ple valuate our PeTTA and live other compara TTA mothods: LEGISLATION TTA ^{s in version} and the conjunction use of Fig. 1 for Section group of finite We evaluate our PeTTA and five other comparable TTA methods in recurring TTA setting on ImâgeÑet-C dataset

does not address the regularization coefficients

 \Box

of regularizer $\mathcal{R}(\theta)$ egularization coefficients
of regularizer $\mathcal{R}(\theta)$ does not address the set of Ω To maintain persistence, utilizing all research and components is suggested in PeTTA $s = \frac{1}{2}$ conjunction with $\frac{1}{2}$ coefficient. egradation **components is suggested in** components is suggested in PeTTA

Ablation Studies: PeTTA sho similarity in conjunction with Fisher \mathcal{Z} and \mathcal{Z} and \mathcal{Z} are conjunction with \mathcal{Z} and \mathcal{Z}

adaptive update the update the update the update of the update the set of the set of the set of the set of the s
The maintain set of the set of t sung performance across zu recurring TTA visits performance. Existing studies fail to detect this issue since PeTTA shows a *persisting performance* across 20 recurring TTA visits

Com PeTTA favors various choices **of Without us** of regularizer $R(\theta)$

Without using/ fixed

the lowest

Conclusions: Persistent Test-time Adaptation (PeTTA)

Conducting theoretical analysis on performance degradation of TTA on ϵ −GMMC, indicating factors that contribute to model collapse

Introducing a new testing scenario *– recurring TTA* for demonstrating the performance degradation of existing continual TTA methods

For more information, visit our project page at \bullet See you at: *POSTER SECTION 4 (Thursday Afternoon)*

Introducing a new baseline – *persistent TTA (PeTTA).* PeTTA strikes a balance between two objectives: adaptation and collapse prevention