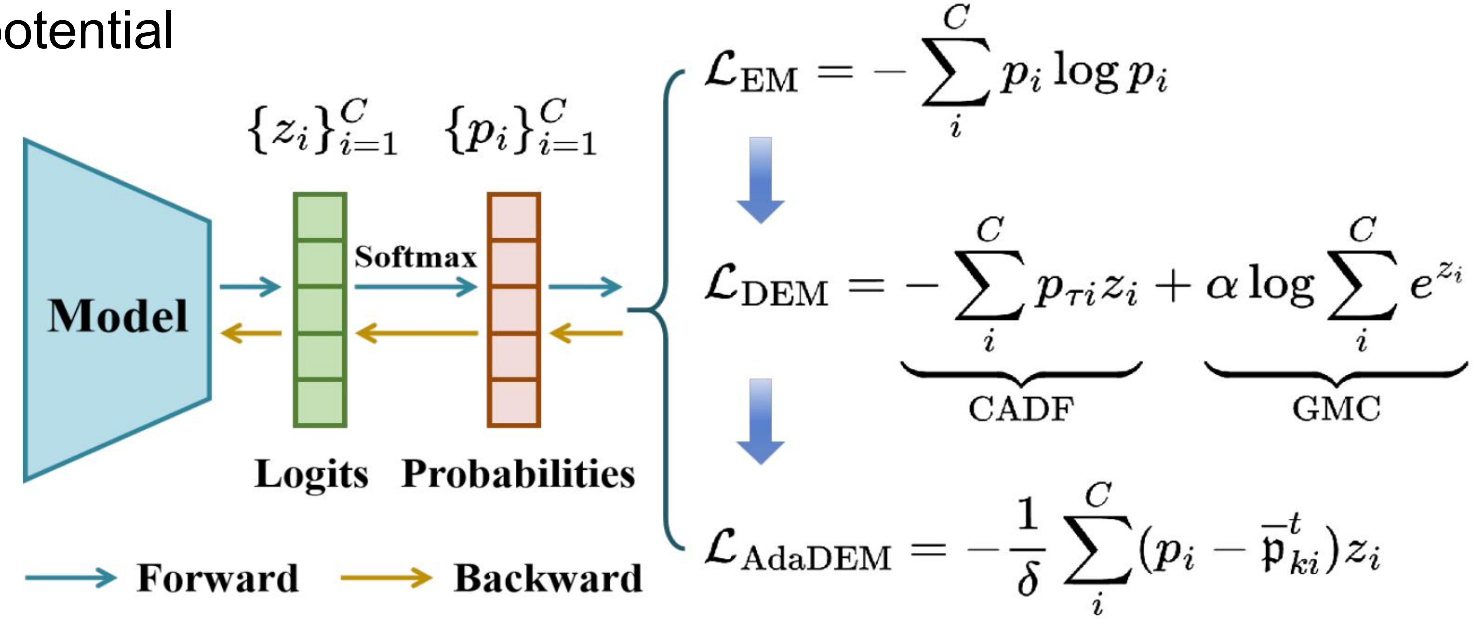
Decoupled Entropy Minimization

Jing Ma, Hanlin Li, Xiang Xiang

Introduction

➢ Entropy Minimization (EM)

- **Confident** & **Accurate** Predictions
- “Sharper” outputs and High-confidence predictions
- "**Coupled**" mechanism limits its potential
  - <u>**Reward Collapse**</u>
  - <u>**Easy-Class Bias**</u>



<u>**AdaDEM: Adaptive Decoupled Entropy Minimization**</u>

Introduction

➢ Decoupled EM (DEM)

- EM is an "**coupling**" of two opposing forces:
- **Cluster Aggregation Driving Factor (CADF)**:

  It rewards classes with higher confidence, pushing the model's prediction to “collapse” toward one extreme.
- **Gradient Mitigation Calibrator (GMC)**:

  It penalizes classes that are already overconfident, preventing the model from becoming too “self-assured”.

$$H(\mathbf{z}) = -\sum_{i=1}^{C} p_i(\mathbf{z}) \log p_i(\mathbf{z})$$

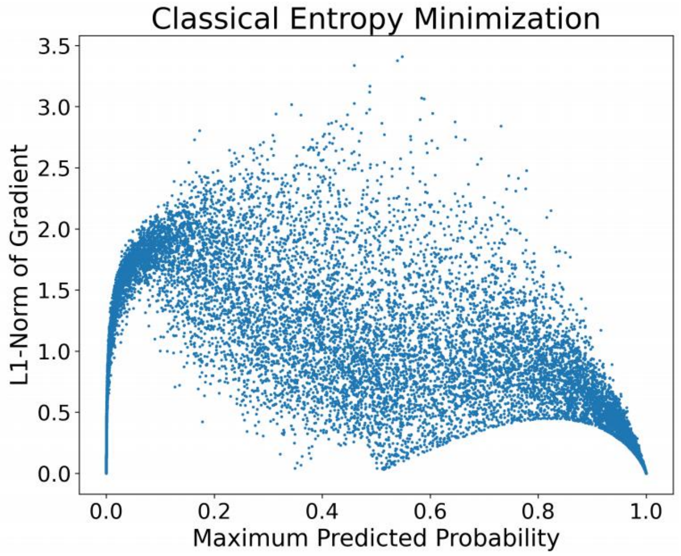
$$H(\mathbf{z}) = -\sum_{i=1}^{C} p_i(\mathbf{z}) \log \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} = \underbrace{-\sum_{i=1}^{C} p_i(\mathbf{z}) z_i}_{\text{CADF}} + \underbrace{\log \sum_{i=1}^{C} e^{z_i}}_{\text{GMC}}$$

Introduction

➢ The Flaws of EM

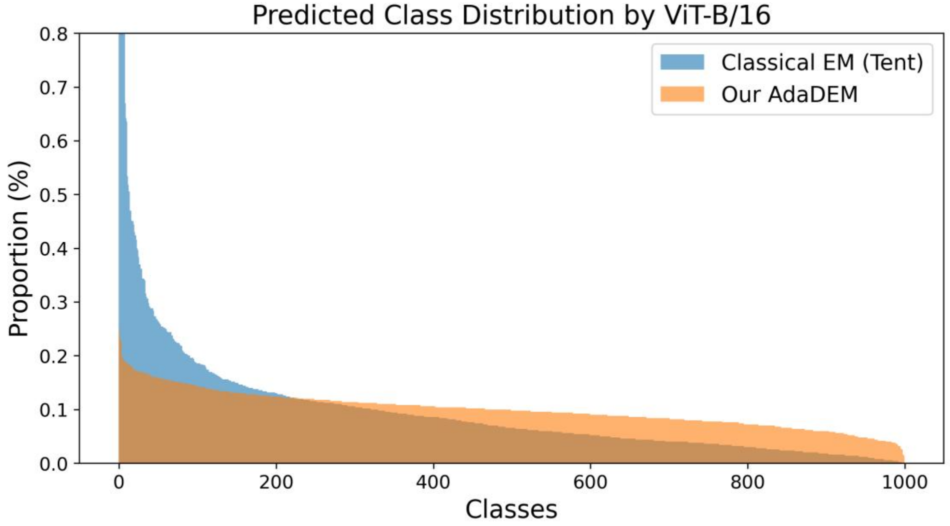
Reward Collapse:

- The samples the model has already "learned" contribute almost no further force during subsequent optimization, which limits their potential for additional performance gains.



Easy-Class Bias:

- This causes the model's predicted distribution to deviate from the true label distribution, showing a particular bias towards classes that are easier to learn.



AdaDEM

Normalized Rewards to Overcome "Reward Collapse":

- Normalizing the rewards generated by the CADF using the L1 norm.
- It ensures that regardless of how high a sample's initial confidence is, the **normalized reward signal** remains at a stable magnitude.
- This guarantees that even high-confidence samples continue to contribute to the model's optimization, thereby breaking through the performance bottleneck.

$$H(\mathbf{z}) = \boxed{-\frac{1}{\delta}} \sum_{i=1}^{C} (p_i(\mathbf{z}) - \overline{\mathfrak{p}}^{t}_{ki}) z_i \Longrightarrow \delta = \| -\partial T(\mathbf{z}|x,\theta)/\partial \mathbf{z} \|_1$$

The forum of AdaDEM

T is the CADF

z is the model's output logits

AdaDEM

Marginal Entropy Calibrator (MEC) to Mitigate "Easy-Class Bias":

- A designed Marginal Entropy Calibrator (MEC) is used to replace the original GMC.
- Unlike GMC only focuses on the prediction of a single sample, MEC takes a global perspective. It estimates the marginal class distribution and uses this estimation to calibrate the gradients.
- This allows the model to **reduce its excessive preference** for certain easy-to-learn classes and ensuring the final output distribution aligns more closely with the true label distribution.

$$H(\mathbf{z}) = -\frac{1}{\delta}\sum_{i=1}^{C}(p_i(\mathbf{z}) - \boxed{\overline{\mathfrak{p}}_{ki}^{t}})z_i \quad \Rightarrow \quad \overline{\mathfrak{p}}_k^t = 0.9 \cdot \overline{\mathfrak{p}}_k^{t-1} + \boxed{0.1/N_k \cdot \sum^{N_k} \mathfrak{p}_k^t}$$

The forum of AdaDEM

t is the iteration index

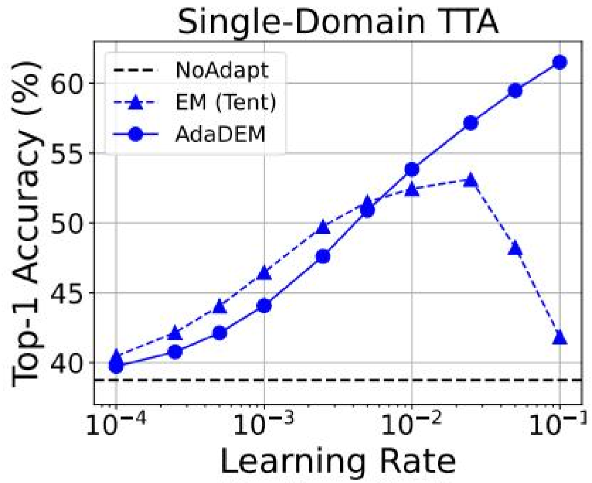
Experiments

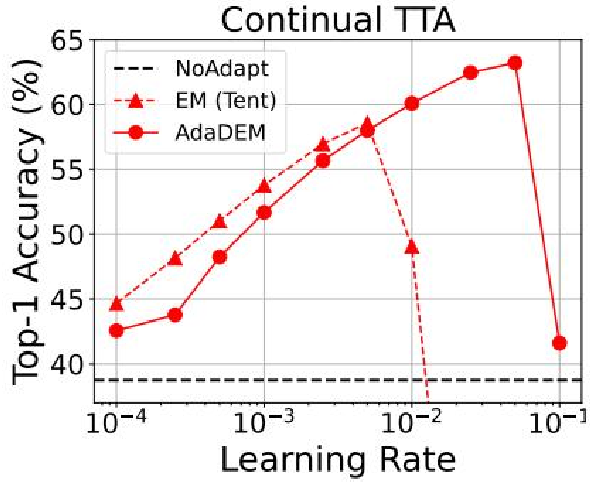
In Single-domain & Continual **Test-Time Adaptation (TTA)** tasks, AdaDEM delivered outstanding performance, **outperforming the EM and its upper-bound variant, DEM*,** significantly less sensitive to the learning rate.

Table 2: Experiments on single-domain & continual TTA tasks (left) and the test-time prompt tuning task (right). Top-1 classification accuracy (%) is reported. We highlight the highest accuracy in **bold** and the second best as underline. Δ denotes the performance improvement relative to the baselines.

Methods	Single-Domain Mean	Single-Domain Δ	Continual Mean	Continual Δ
NoAdapt	38.8±0.00	-	38.8±0.00	-
Tent†	53.1±0.65	+0.0	58.6±0.09	+0.0
+ DEM*	56.0±0.32	+2.9	64.1±0.05	+5.5
+ AdaDEM	61.5±0.20	+8.4	63.2±0.16	+4.6
Tent	52.7±0.10	+0.0	48.5±0.71	+0.0
+ DEM*	55.1±0.11	+2.4	64.5±0.14	**+16.0**
+ AdaDEM	66.2±0.12	**+13.5**	64.4±0.02	+15.9
ETA	65.1±0.10	+0.0	64.2±0.04	+0.0
+ DEM*	66.3±0.04	+1.2	65.7±0.04	+1.5
+ AdaDEM	**66.8**±0.02	+1.7	66.1±0.01	+1.9
EATA	62.2±0.14	+0.0	64.9±0.08	+0.0
+ DEM*	64.4±0.30	+2.2	66.2±0.07	+1.3
+ AdaDEM	65.3±0.11	+3.1	**66.4**±0.04	+1.5
DeYO	62.6±0.32	+0.0	57.6±0.36	+0.0
+ DEM*	65.6±0.03	+3.0	65.4±0.12	+7.8
+ AdaDEM	62.6±0.10	+0.0	59.0±0.05	+1.4
SAR	54.2±0.07	+0.0	57.0±0.05	+0.0
+ DEM*	57.9±0.04	+3.7	62.4±0.03	+5.4
+ AdaDEM	65.7±0.07	+11.5	63.0±0.05	+6.0

Methods	ImageNet -1K	ImageNet -A	ImageNet -V2.	ImageNet -R.	ImageNet -S.	Avg.	Δ
CLIP-RN50							
Zero-Shot	58.2±0.00	21.8±0.00	51.4±0.00	56.2±0.00	33.4±0.00	44.2±0.00	+0.0
Ensemble	59.8±0.00	23.2±0.00	52.9±0.00	**60.7**±0.00	35.5±0.00	46.4±0.00	+2.2
TPT	60.7±0.07	26.1±0.10	54.6±0.02	58.9±0.08	35.2±0.09	47.1±0.06	+2.9
+ DEM*	61.3±0.09	25.5±0.07	55.0±0.10	59.7±0.12	35.6±0.08	47.4±0.04	+3.2
+ AdaDEM	60.7±0.04	29.2±0.19	54.8±0.22	58.8±0.05	35.4±0.03	47.8±0.07	+3.6
CoOp	63.3±0.00	23.1±0.00	55.4±0.00	56.6±0.00	34.7±0.00	46.6±0.00	+2.4
TPT (CoOp)	65.4±0.06	28.9±0.14	58.2±0.10	59.0±0.09	36.3±0.15	49.6±0.07	+5.4
+ AdaDEM	**65.6**±0.05	**31.3**±0.10	**58.5**±0.22	59.3±0.10	**36.3**±0.11	**50.2**±0.06	**+6.0**
CLIP-ViT-B/16							
Zero-Shot	66.7±0.00	47.9±0.00	60.9±0.00	74.0±0.00	46.1±0.00	59.1±0.00	+0.0
Ensemble	68.3±0.00	49.9±0.00	61.9±0.00	77.7±0.00	48.2±0.00	61.2±0.00	+2.1
TPT	69.0±0.04	54.5±0.09	63.4±0.13	77.0±0.06	48.0±0.13	62.4±0.05	+3.3
+ DEM*	68.9±0.03	54.8±0.09	63.5±0.11	77.1±0.08	47.9±0.06	62.5±0.06	+3.4
+ AdaDEM	69.4±0.12	58.8±0.18	64.0±0.06	77.6±0.21	48.6±0.05	63.7±0.05	+4.6
CoOp	71.5±0.00	49.7±0.00	64.2±0.00	75.2±0.00	48.0±0.00	61.7±0.00	+2.6
TPT (CoOp)	73.6±0.05	57.9±0.12	66.9±0.08	77.2±0.04	49.2±0.07	64.9±0.06	+5.8
+ AdaDEM	**73.7**±0.07	**60.3**±0.11	**66.9**±0.19	**77.9**±0.14	**49.3**±0.07	**65.6**±0.02	**+6.5**





Experiments

➢ Semi-Supervised Learning (SSL)

Table 3: Experiments on semi-supervised learning tasks. We report the average Top-1 classification accuracy (%) under different numbers of labeled samples. We highlight the highest accuracy in **bold** and the second best as underline. Δ denotes the performance improvement relative to the baselines.

Methods	CIFAR-10 Mean	CIFAR-10 Δ	CIFAR-100 Mean	CIFAR-100 Δ	STL-10 Mean	STL-10 Δ	EuroSat Mean	EuroSat Δ	TissueMNIST Mean	TissueMNIST Δ	Semi-Aves Mean	Semi-Aves Δ
Ent. Min.	96.4±1.9	+0.0	72.6±0.5	+0.0	82.4±1.6	+0.0	76.5±3.6	+0.0	47.3±2.6	+0.0	59.9±0.7	+0.0
+ AdaDEM	97.2±0.2	+0.8	75.8±0.3	**+3.2**	84.8±0.3	**+2.4**	83.7±0.8	**+7.2**	49.3±1.3	+2.0	61.0±0.2	**+1.1**
Vat (w/ Ent. Min.)	97.4±2.0	+0.0	77.5±0.7	+0.0	85.4±0.9	+0.0	86.6±5.4	+0.0	45.2±4.8	+0.0	61.0±0.4	+0.0
+ AdaDEM	98.5±0.1	**+1.1**	78.8±0.7	+1.3	86.9±0.1	+1.5	91.2±0.9	+4.6	49.4±0.6	**+4.2**	61.8±0.0	+0.8
MixMatch	98.5±0.3	+0.0	73.8±0.4	+0.0	82.9±2.1	+0.0	79.3±4.3	+0.0	48.0±1.7	+0.0	62.6±0.2	+0.0
+ AdaDEM	98.3±0.6	-0.2	74.8±0.2	+1.0	85.2±0.1	+2.3	83.9±2.4	+4.6	**50.0**±0.9	+2.0	62.5±0.1	-0.1
FixMatch	98.4±0.7	+0.0	79.3±0.6	+0.0	88.5±1.2	+0.0	91.9±3.6	+0.0	46.7±3.3	+0.0	**68.2**±0.3	+0.0
+ AdaDEM	98.7±0.2	+0.3	79.5±0.9	+0.2	87.9±0.8	-0.6	**96.9**±0.7	+5.0	49.9±1.3	+3.2	67.9±0.1	-0.3
FreeMatch	98.9±0.0	+0.0	82.6±0.1	+0.0	89.7±1.6	+0.0	95.8±1.0	+0.0	45.6±3.3	+0.0	67.0±0.3	+0.0
+ AdaDEM	**99.0**±0.1	+0.1	**83.1**±0.1	+0.5	**91.6**±0.2	+1.9	95.9±0.1	+0.1	47.9±0.4	+2.3	67.3±0.2	+0.3

➢ Unsupervised Domain Adaptation (UDA) for Semantic Segmentation

Table 4: Experiments on unsupervised domain adaptation task of semantic segmentation. Standard mIoU (%) is reported.

Methods	MinEnt	+ AdaDEM	AdvEnt	+ DEM*	+ AdaDEM
mIoU	41.6±0.47	42.7±0.13	43.6±0.19	44.6±0.32	**44.9**±0.17
Δ	+0.0	+1.1	+0.0	+1.0	**+1.3**

Experiments

➢ **Class-Imbalanced / Long-Tail (LT) Classification**

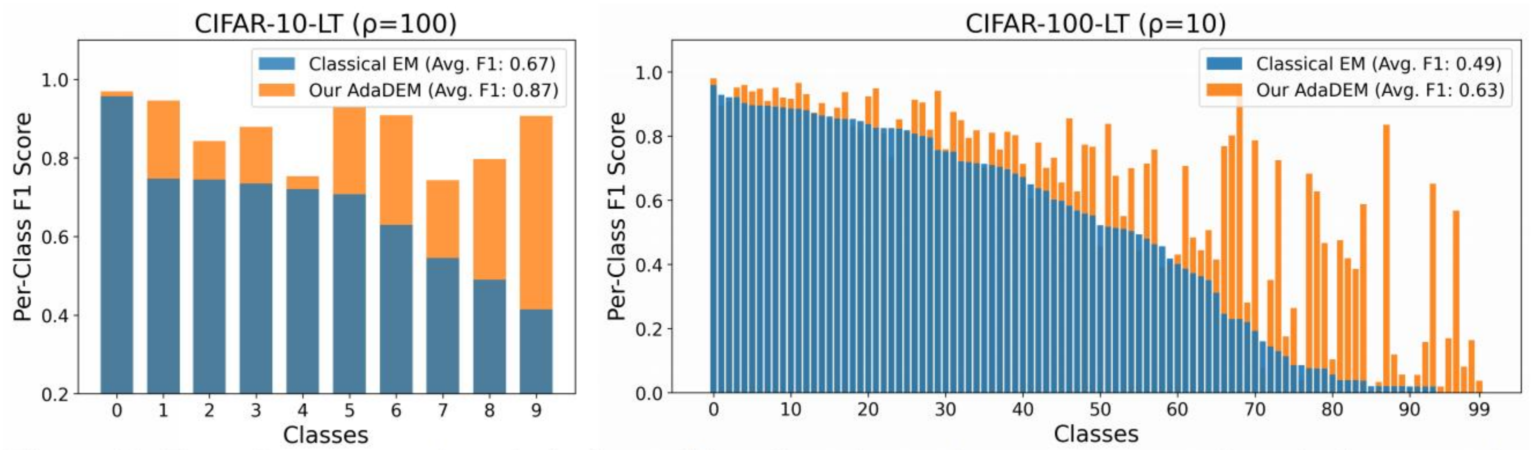


Figure 5: Experiments on class-imbalanced benchmarks. ρ denotes the sample ratio between the most and least populous classes. Both per-class F1 scores and the average F1 score are reported.

➢ **Reinforcement Learning (RL)**

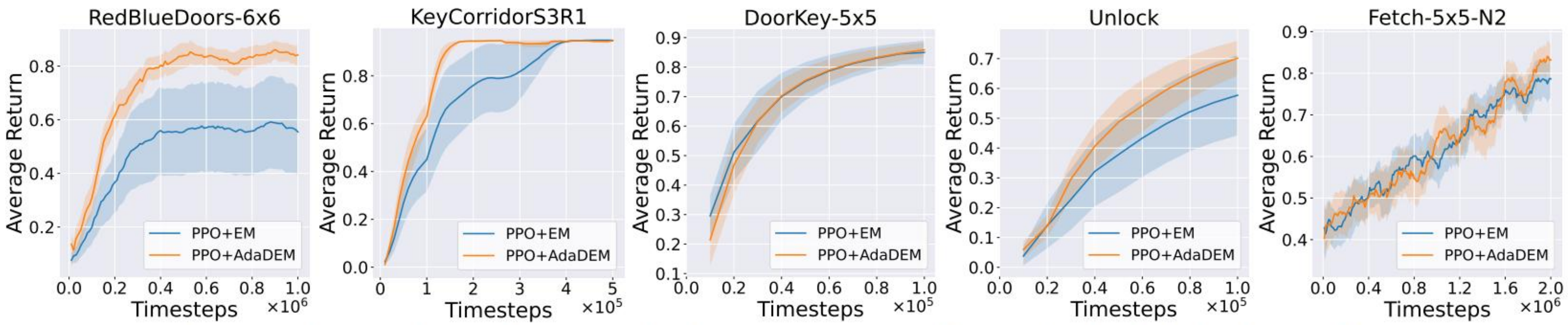


Figure 6: Experimental results on Reinforcement Learning tasks in Minigrid environments.

Thank You

Get the performance upgrade with just one line of code.

More interesting analyses in the paper !

Code available on Github !