BMVC 2022

The 33rd British Machine Vision Conference

Adaptive-TTA: accuracy-consistent weighted test time augmentation method for the uncertainty calibration of deep learning classifiers

Pedro Conde*, Cristiano Premebida*

pedro.conde@isr.uc.pt

cpremebida@isr.uc.pt

*University of Coimbra, Institute of Systems and Robotics,
Department of Electrical and Computer Engineering, Coimbra, Portugal

Introduction

Building deep machine learning systems to classify image data in real-world applications requires not only a quantification of the accuracy of the models but also an understanding of their reliability. With this in mind, the **uncertainty calibration of Deep Neural Networks** in the task of image classification is addressed in this work. We propose a novel technique based on test time augmentation - ***Adaptive-TTA*** - that, unlike traditional test time augmentation approaches, improves uncertainty calibration without affecting the model's accuracy, by leveraging an adaptive weighting system.

The Problem of Uncertainty Calibration

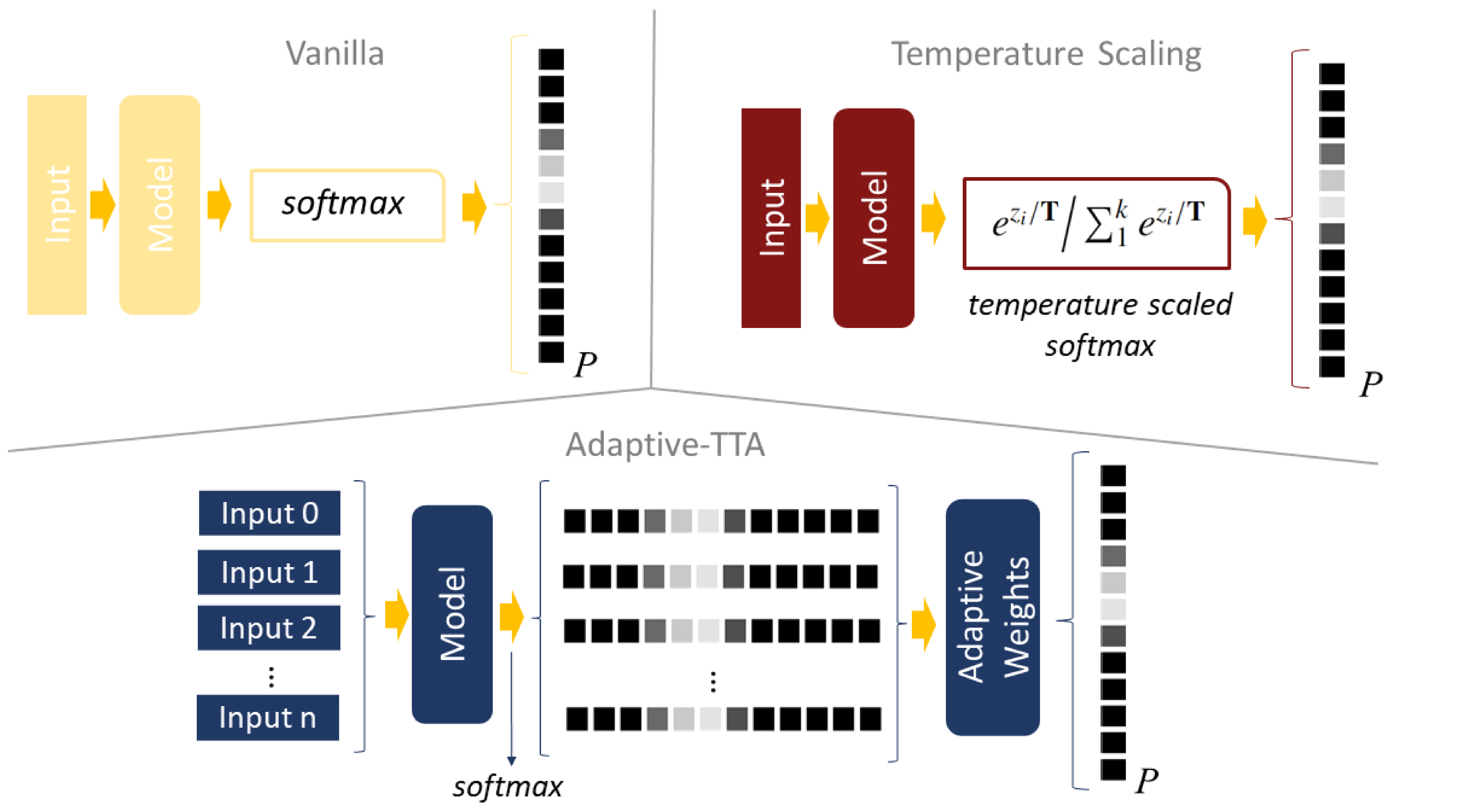
Let us consider X an input space, Y the corresponding set of true labels and a model $f: X \to \Delta_k$, with $\Delta_k = \{(p_1, \dots, p_k) \in [0,1]^k : \sum_{i=1}^{k} p_i = 1\}$ a probability simplex. The model f is considered ***calibrated*** if

$$\mathbb{P}[Y = \arg\max_{i\in\{1,\dots,k\}} f(X) \mid \max_{i\in\{1,\dots,k\}} f(X)] = \max_{i\in\{1,\dots,k\}} f(X).$$

Achieving perfect calibration is impossible in practical settings. Furthermore, the probability values in the left hand side of the previous equation cannot be computed using finitely many samples, which motivates the need for scoring rules to assess uncertainty calibration like the *Brier score.*
For a set of N predictions we define the ***Brier score*** (↓) as

$$BS = \frac{1}{N}\sum_{j=1}^{N}(p^j - o^j)^2,$$

where p^j is the highest confidence value of the prediction j and o^j equals 1 if the true class corresponds to the prediction, and 0 otherwise.



Adaptive-TTA

For m different type of augmentations each one applied n_i times ($i = 1, \dots, m$), and for some vector $(\omega_1, \omega_2, \dots, \omega_m) \in \mathbb{R}^m$ and some value $\omega^* \in [0,1]$ (see *Experiments*), we define our prediction probability vector as

$$\mathbf{p}(\bar{\omega}) = (1-\bar{\omega})\mathbf{p}_0 + \frac{\bar{\omega}\sum_{i=1}^{m}|\omega_i|\sum_{j=1}^{n_i}\mathbf{p}_j^i/n_i}{\sum_{i=1}^{m}|\omega_i|}$$

with

$$\bar{\omega} = \max\left\{\omega \in [0,\omega^*] : \arg\max_{i\in\{1,\dots,k\}} \mathbf{p}(\omega) = \arg\max_{i\in\{1,\dots,k\}} \mathbf{p}_0\right\}.$$

The vector $\mathbf{p}_j^i$ denotes the probability prediction vector associated with the j-nth augmentation of the i-nth type. Also, k refers to the number of classes.
The value of $\bar{\omega}$ may vary in each prediction, adapting in a way that prevents corruptions in terms of accuracy. In a practical scenario, the value $\bar{\omega}$ is determined in the following way: starting with $\omega^0 \coloneqq \omega^*$, iterating with $\omega^t = \omega^{t-1} - \epsilon$ (in this case $\epsilon = 0.01$) and stopping at the moment t^* when the condition

$$\arg\max_{i\in\{1,\dots,k\}} \mathbf{p}\left(\omega^{t^*}\right) = \arg\max_{i\in\{1,\dots,k\}} \mathbf{p}_0$$

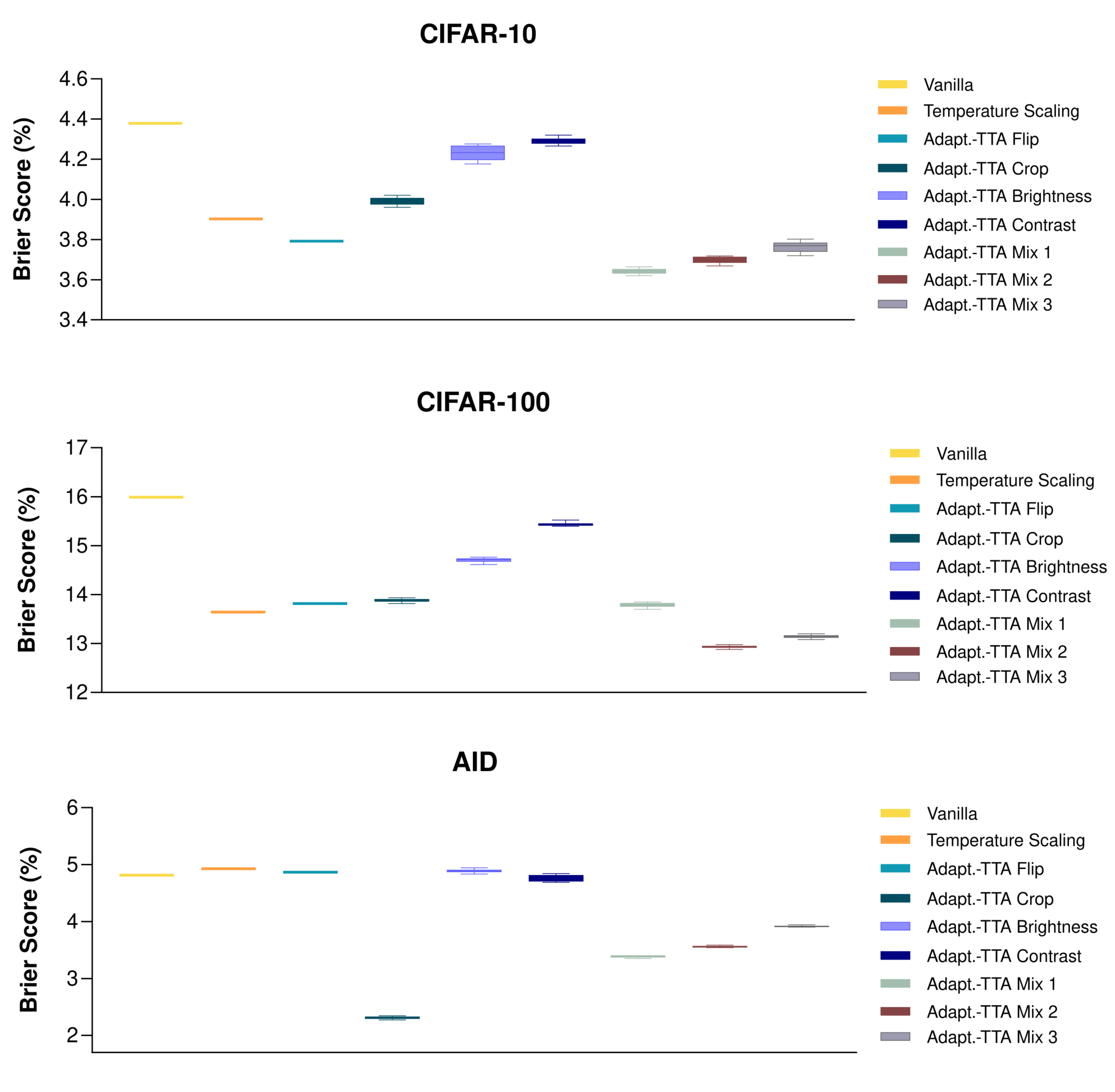
is satisfied, thus defining $\bar{\omega} \coloneqq \omega^{t^*}$.

Experiments

For each experiment done with ***Adaptive-TTA***, the parameters $\omega^* \in [0,1]$ and $(\omega_1, \omega_2, \dots, \omega_m) \in \mathbb{R}^m$ are optimized on a given validation set, using the Expected Calibration Error – with 15 bins – as loss function.

Adaptive-TTA was applied with seven different augmentation policies. *Flip* consists of one flip transformation (around the vertical axis); *Crop* consists of five crop transformations with 78% of the original size, in random position; *Brightness* consists of five brightness transformations with a random intensity value in the interval [-0.5,0.5]; *Contrast* consists of five contrast transformations with a random intensity value in the interval [-0.2,0.2]; *Mix 1* combines the augmentations present *Flip* and *Crop*; *Mix 2* combines the augmentations present in *Mix 1* and *Brightness*; *Mix 3* combines the augmentations present in *Mix 2* and *Contrast*. Given the randomness inherent to some transformation parameters, some results are presented in the form of box plots, resulting from 10 different experiments.

Results



Acknowledgments

This work has been supported by the Portuguese Foundation for Science and Technology (FCT), via the project *GreenBotics* (PTDC/EEI-ROB/2459/2021).