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

Pedro Conde
pedro.conde@isr.uc.pt
Cristiano Premebida
cpremebida@isr.uc.pt

University of Coimbra
Institute of Systems and Robotics
Coimbra, PT

Abstract

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, called *Adaptive-TTA*, that - unlike traditional test time augmentation approaches - improves uncertainty calibration without affecting the model's accuracy. This technique is evaluated with respect to the *Brier score* - a proper scoring rule for measuring the calibration of predicted probabilities - on the classical CIFAR-10/CIFAR-100 computer vision datasets, as well as on the benchmark satellite imagery dataset AID, using different augmentation policies. Our approach outperforms *temperature scaling*, a state-of-the-art *post-hoc* calibration technique, on all the three aforementioned datasets.

1 Introduction

Real-world applications of machine learning (ML) systems require a thorough look into the reliability of the learning models and consequently to their uncertainty calibration (also referred as confidence calibration or simply *calibration*). In addition to having highly accurate classification models, the user should be able to "trust" their predictions, specially when dealing with critical application domains[1], where wrong decisions can result in potentially dramatic consequences. To do so, it is required that the confidence output generated by the referred ML models (that translates the confidence the model has in the prediction that is making) realistically represents the correct likelihood of its prediction - *i.e.*, the model is *calibrated*. A calibrated model allows for an accurate quantification of predictive uncertainty, which results in reliable confidence values associated with its prediction.

Although successful ML approaches have been proposed in recent years in classification tasks for a multitude of applications, due to the accuracy of modern Deep Neural Networks

[1] For example, self-driving cars and autonomous robotics.

(DNNs), such deep models have been found to be tendentiously *uncalibrated* [6, 23], making either overconfident or underconfident predictions, which make them unusable in scenarios where wrong classifications may carry undesirable consequences. Therefore, improving the calibration of these Deep Learning (DL) models is the goal of the method presented in this work, which is based on a novel approach to test time augmentation.

Test time augmentation is a methodology that relies on data augmentation techniques to create multiple samples from the original input at inference. Contrarily to traditional data augmentation methodologies, the proposed augmentations are, in this case, carried out only before prediction (at the testing phase) and not during the training process of the machine learning models, therefore being easily applied to pre-trained models. Inspired in this type of techniques, we introduce *Adaptive-TTA* that leverages the use of test time augmentation combined with a custom weighting system – that guaranties consistency in terms of the model's accuracy – to obtain better calibrated outputs in DNNs. Our experiments are made in the classical CIFAR-10/CIFAR-100 [12] datatasets, as well as in a benchmark satellite imagery dataset, with 30 different classes, named Aerial Image Dataset (AID) [31]. The results are compared with the performance of *temperature scaling*, a state-of-the-art *post-hoc* calibration method, with respect to the *Brier score* [2].

Our contribution is twofold:

- A novel technique based on test time augmentation, named *Adaptive-TTA*, that guarantees consistency in the model's accuracy while improving its uncertainty calibration;
- A study of the effects of different augmentation policies in the uncertainty calibration of the presented models, in the context of the novel technique introduced.

2 Related Work

The topic of uncertainty calibration in DNNs has been introduced in [6], where the authors evaluate calibration in various datasets in both Computer Vision and Natural Language Processing applications, with different modern DNNs. They argued that although more recent DL architectures have allowed for improved accuracy in various tasks, modern DL models are often less calibrated than older counterparts.

One popular calibration metric is the *Expected Calibration Error* (ECE) [19], that evaluates the bin-wise difference between accuracy and confidence. Nevertheless, some authors have identified limitations regarding the usage of ECE as a calibration metric [7, 22, 30], namely its dependence on the selected binning scheme and intractability. Additionally, ECE is not a proper scoring rule [4] like for example *Brier score* [2] - an increasingly popular metric to assess the calibration of predicted probabilities [11, 15, 23, 27] - that overcomes some of the limitations of ECE.

Post-hoc calibration methods - which can be applied after the training process of a DNN - are used in [6] to tackle the referred calibration problems. *Temperature scaling* is introduced as an extension of the *Platt scaling* algorithm [21, 25], and has the best performance on most datasets against other algorithms like histogram binning [32] and isotonic regression [33]. Given its good performance, *temperature scaling* is usually used as baseline for comparison of calibration methods [2, 9, 11, 13, 23].

Other approaches, like approximate Bayesian models [3, 6] and some regularization techniques [17, 24], have also been used in the context of uncertainty calibration [23]. The

caveats are that these approaches require building new, and more complex models, or modifying and re-training pre-existing ones, contrarily to the previously mentioned *post-hoc* calibration methods and the method proposed in this work.

In recent years, some attention has been given to test time augmentation methods - especially in biomedical applications [16, 28, 29] - although, to the best of our knowledge, all relevant literature fails to address its effect on calibration-specific metrics like the *Brier score*, or even ECE. The use of this technique opens the possibility for a multitude of augmentation policies; for instance, both [10] and [14] focus on learnable augmentation policies, which falls out of the scope of this work, but may open different possibilities for the application of methods based on test time augmentation. It is worthy of note that the work in [26] makes an interesting study on the effects of test time augmentation on the accuracy of models, showing that such technique may produce corrupted predictions which can ultimately worsen the model's performance. This is one the motivations of the work here developed, since our novel approach allows the usage of test time augmentation in the context of uncertainty calibration without corrupting the model's accuracy.

3 Background

Notation For the remainder of this work we will use bold notation to denote vectors, like $\mathbf{p} = (p_1, \ldots, p_k)$. We can also refer to the i-nth element of the vector $\mathbf{p}$ as $\mathbf{p}_{(i)} := p_i$. Finally, the $\downarrow$ symbol, associated with a given metric, informs that a lower value of such metric represents a better performance.

In this section we describe practical metrics for assessing uncertainty calibration, as well as the popular *post-hoc* calibration method - *temperature scaling* - that will be used as baseline for comparison against the approach proposed in this work. Besides *Brier score* - that will serve to evaluate the performance of our method - we will also describe ECE, since it has been empirically found useful in the parameter optimization of *Adaptive-TTA*. Contrarily to *Brier score*, ECE is not a proper scoring rule and so there exist trivial uninformative solutions that result in an optimal score (e.g., always returning the marginal probability), thus making it less useful when evaluating the predictive uncertainty of a given model.

We will start by making a brief introduction to the concept 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 $\Delta_k = \{(p_1, \ldots, p_k) \in [0,1]^k : \sum_{i=1}^{k} p_i = 1\}$ a probability simplex. For some $x \in X$, the corresponding label $y \in Y$, $\mathbf{p} = (p_1, \ldots, p_k) \in \Delta_k$ and

$$c(x) = \underset{i \in \{1,\ldots,k\}}{\arg\max}\, f(x), \tag{1}$$

a predicted confidence value p_i is considered calibrated if

$$p_i = P(y = i | f(x) = \mathbf{p}, c(x) = i). \tag{2}$$

The model f is considered calibrated if it only outputs calibrated predictions. However, as stated in [6], achieving perfect calibration is impossible in practical settings. Furthermore, the probability in the right hand side of (2) cannot be computed using finitely many samples, which motivates the need for scoring rules to assess uncertainty calibration.

3.1 Brier score ↓

Brier score [2] is a proper scoring rule [9] that computes the squared error between a predicted probability and its true response, hence its utility to evaluate model calibration. For a set of N predictions we define the *Brier score* as

$$BS = \frac{1}{N}\sum_{j=1}^{N}(p^j - o^j)^2, \tag{3}$$

where p^j is the confidence value of the predicted class and o^j equals 1 if the true class corresponds to the prediction and 0 otherwise (we use here superscript indexation to avoid confusion with the previous notation). We refer to [18] and [1] for some thorough insights about the interpretability and decomposition of the *Brier score*.

3.2 Expected Calibration Error ↓

To compute the *Expected Calibration Error* (ECE) [19] we start by dividing the interval $[0,1]$ in M equally spaced intervals. Then a set of bins $\{B_1, B_2, \ldots, B_M\}$ is created, by assigning each predicted probability value to the respective interval. The idea behind this measurement is to compute a weighted average of the absolute difference between accuracy and confidence in each bin B_i ($i = 1, \ldots, M$). We define the confidence for each bin as

$$conf(B_i) = \frac{1}{|B_i|}\sum_{j \in B_i} p^j, \tag{4}$$

where p^j is the predicted confidence for the sample j, and accuracy as

$$acc(B_i) = \frac{1}{|B_i|}\sum_{j \in B_i} o^j, \tag{5}$$

where o^j is defined as previously in (3). Then the ECE is calculated, for a total of N samples, as

$$\mathrm{ECE} = \sum_{i=1}^{M}\frac{|B_i|}{N}|conf(B_i) - acc(B_i)|. \tag{6}$$

3.3 Temperature scaling

Temperature scaling is the most popular *post-hoc* calibration technique and it has been shown to be a robust method in various applications [6]. As referred before, this method can be used with models that are already trained and ready to be deployed, contrarily to other techniques that require re-training, modifying or rebuilding the models.

For a classification problem with k different classes and for a logit vector $\mathbf{z} = (z_1, \ldots, z_k)$, we replace the usual *softmax* function

$$\sigma_{SM}(\mathbf{z})_{(i)} = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}, \qquad i = 1, 2, \ldots, k, \tag{7}$$

with the *temperature scaled* version, for $T > 0$,

$$\sigma_{TS}(\mathbf{z})_{(i)} = \frac{e^{z_i/T}}{\sum_{j=1}^{k} e^{z_j/T}}, \qquad i = 1, 2, \ldots, k. \tag{8}$$

This parameter T is optimized w.r.t. the Negative Log Likelihood (NLL) in a validation set. Since T is the same for all classes (and $T > 0$), the following equality holds true

$$\max_{i \in \{1,\ldots,k\}} \sigma_{SM}(\mathbf{z}) = \max_{i \in \{1,\ldots,k\}} \sigma_{TS}(\mathbf{z}), \tag{9}$$

which means that temperature scaling will not change the model's accuracy.

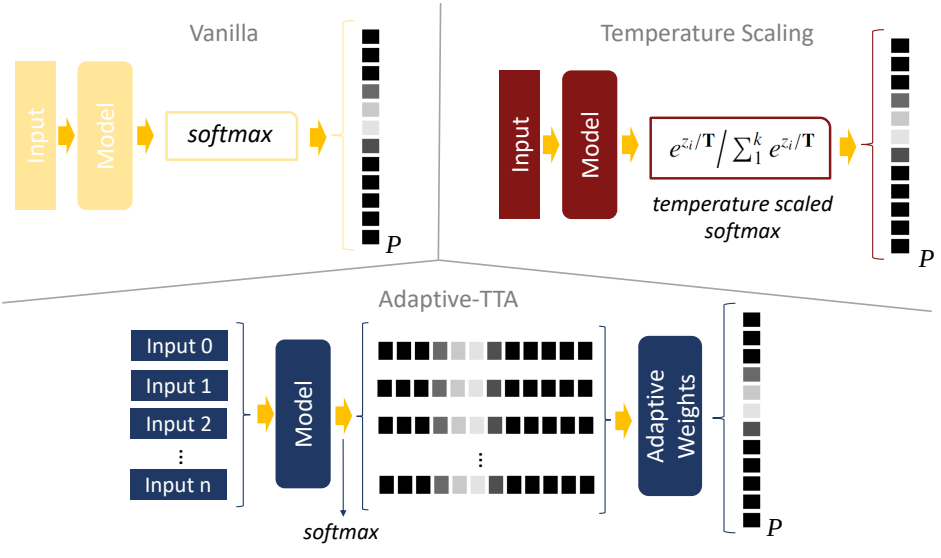

Figure 1: Comparative scheme illustrating, from a "high level" perspective, the general differences between the approaches evaluated in this work - *vanilla* (referring to the results obtained from the DNN without any type of calibration method), *temperature scaling* and our novel method *Adaptive-TTA*.

4 Proposed Method: Adaptive-TTA

In this section we describe the proposed calibration method - *Adaptive-TTA* - that consists in an innovative approach to test time augmentation. We start by introducing a classical naive approach to test time augmentation and then detail the differences in our methodology.

Test time augmentation leverages the use of data augmentations only before inference, contrarily to traditional augmentation methods that are applied during the training process of a given model. Thus, just like *temperature scaling* (previously described), this methodology is easily applied to pre-trained models. Formally, for an input I_0, we conduct n different transformations, obtaining $n+1$ different inputs $I_0, I_1, \ldots, I_n$; then, we feed all the different

inputs to our model resulting in $n+1$ different prediction probability vectors $\mathbf{p}_0, \mathbf{p}_1, \ldots, \mathbf{p}_n$; finally all the predictions are averaged to obtain the final prediction probability vector

$$\mathbf{p} = \frac{\sum_{i=0}^{n} \mathbf{p}_i}{n+1}. \tag{10}$$

A naive application of test time augmentation, like in (10), can corrupt predictions and ultimately have a negative impact on the model's accuracy [26]. Hence, we propose *Adaptive-TTA* that, in line with *temperature scaling*, does not alter the class predicted by the model in which is applied and therefore can be specifically optimized to tackle problems regarding uncertainty calibration, without the concern of corrupting the model's accuracy in the process. Let us note that, given that most common augmentations have parameters with random properties, it can be useful to apply the same type of augmentation more than once, because it produces different results. For the same reason (randomness in parameters) it only makes sense to find one weight per augmentation type. As such, in the case of *Adaptive-TTA*, for m different types of augmentation, each one applied n_i times ($i = 1, \ldots, m$), and for some vector $(\omega_1, \omega_2, \ldots, \omega_m) \in \mathbb{R}^m$, we define our prediction probability vector as

$$\mathbf{p}(\bar{\omega}) = (1-\bar{\omega})\mathbf{p}_0 + \frac{\bar{\omega} \sum_{i=1}^{m} |\omega_i| \sum_{j=1}^{n_i} \mathbf{p}_j^i / n_i}{\sum_{i=1}^{m} |\omega_i|} \tag{11}$$

with

$$\bar{\omega} = \max \left\{ \omega \in [0, \omega^*] : \underset{i \in \{1,\ldots,k\}}{\arg\max}\, \mathbf{p}(\omega) = \underset{i \in \{1,\ldots,k\}}{\arg\max}\, \mathbf{p}_0 \right\}. \tag{12}$$

For the sake of clarity, let us note that 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, accordingly to the definition in (12). In a practical scenario, the value $\bar{\omega}$ is determined in the following way: starting with $\omega^0 := \omega^*$, iterating with $\omega^t = \omega^{t-1} - \varepsilon$ (in our case $\varepsilon = 0.01$) and stopping at the moment t^* when the condition

$$\underset{i \in \{1,\ldots,k\}}{\arg\max}\, \mathbf{p}\left(\omega^{t^*}\right) = \underset{i \in \{1,\ldots,k\}}{\arg\max}\, \mathbf{p}_0 \tag{13}$$

is satisfied, thus defining $\bar{\omega} := \omega^{t^*}$; this has low computational effort since the probability vectors for each augmentation are only computed in the beginning of the iterations.

Both $\omega^* \in [0,1]$ and $(\omega_1, \omega_2, \ldots, \omega_m) \in \mathbb{R}^m$ are pre-defined, and are dependent on the dataset we are working with; they can be defined by formulating an optimization problem with a give validation set, such as described in Section 5.

Figure 1 shows an illustration of the differences between *Adaptive-TTA* and the other approaches evaluated in this work, from a "high level" perspective.

5 Experiments and Results

In this section we present the results obtained with *Adaptive-TTA* - in terms of *Brier score* - for the three previously referred datasets, CIFAR-10, CIFAR-100 and AID. As mentioned,

the proposed method can be applied with an highly extensive set of augmentations policies, thus, in this work, we present a study of different types of augmentation policies in the context of our novel approach. For practical reasons, the types of transformations used are limited to four: *flip*, *crops* and also changes in *brightness* and *contrast*. A *flip* transformation consists on a flip around the vertical axis of the image input; a *crop* transformation creates a cropped input with dimension of ratio τ from the original input dimension, extracted from a random position within the input image; a *brightness* transformation creates, from an original input γ, a new input

$$\gamma' = \gamma + \beta\gamma_{\max}, \tag{14}$$

where $\gamma_{\max}$ is the maximum pixel value from γ, and β is a random number extracted from a continuous Uniform distribution within a given interval $[\beta_{\min}, \beta_{\max}]$; a *contrast* transformation creates, from an original input γ, a new input

$$\gamma'' = \gamma(1+\alpha), \tag{15}$$

where α is a random number extracted from a continuous Uniform distribution within a given interval $[\alpha_{\min}, \alpha_{\max}]$.

The results obtained using *Adaptive-TTA* are compared against the performance of *temperature scaling* and a *vanilla* approach (referring to the results obtained from the DNN without any type of calibration method). In all experiments, a ResNet [8] architecture is used; the DNNs trained on the CIFAR-10 and CIFAR-100 datasets have 56 residual layers, while the DNN trained with the AID dataset has 50 residual layers. The achieved accuracy values are 94.21%, 72.78% and 93.35%, for CIFAR-10, CIFAR-100 and AID test sets, respectively.

For each experiment done with *Adaptive-TTA*, the parameters

$$\omega^* \in [0,1], \quad (\omega_1, \omega_2, \ldots, \omega_m) \in \mathbb{R}^m, \tag{16}$$

described in Section 4, are optimized on a given validation set (the nature of such validation sets is described in the following subsections). For this optimization we use the ECE, with 15 bins, as the loss function, and a Nelder-Mead optimization algorithm [20]. For the cases where the augmentation policy has parameters with random properties, the loss funtion is the average of the ECE obtained after 10 different experiments. We have empirically found that using ECE as a loss function accomplishes better results than using NLL or the actual *Brier score*, when making this optimization process. We speculate that this choice of loss function prevents the overfitting of the parameters for the given validation set.

5.1 CIFAR-10/CIFAR-100

Both CIFAR-10 and CIFAR-100 datasets are subsets of the *tiny images* dataset [12], with respectively 10 and 100 different equally balanced classes. Both datasets have a total of 60000 RGB images with size 32×32, and are divided in training and test sets, each comprising 50000 and 10000 images, respectively. For the purpose of this work, the first 1000 images of the traditionally defined test set are assigned to a validation set and the remaining 9000 images compose our test set, for both cases.

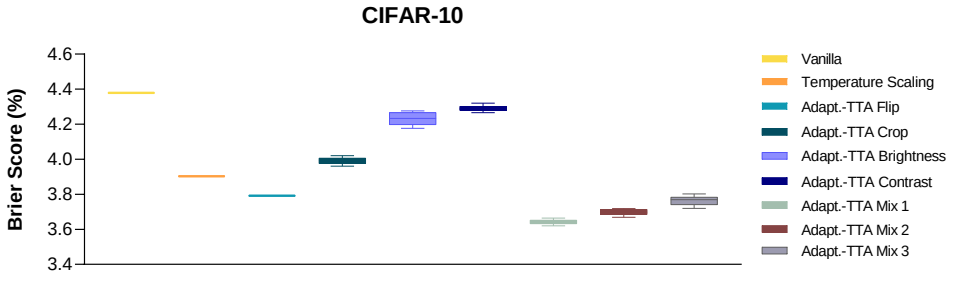

Figure 2: Results on the CIFAR-10 dataset. *Adaptive-TTA Flip* consists of one augmentation of the type *flip*; *Adaptive-TTA Crop*, *Adaptive-TTA Brightness* and *Adaptive-TTA Contrast* consist of five augmentations of the types *crop* ($\tau \approx 0.78$), *brightness* ($\beta_{\min} = -0.5$, $\beta_{\max} = 0.5$) and *contrast* ($\alpha_{\min} = -0.2$, $\alpha_{\max} = 0.2$), respectively; *Adaptive-TTA Mix 1* combines the augmentations present in *Adaptive-TTA Flip* and *Adaptive-TTA Crop*; *Adaptive-TTA Mix 2* combines the augmentations present in *Adaptive-TTA Mix 1* and *Adaptive-TTA Brightness*; *Adaptive-TTA Mix 3* combines the augmentations present in *Adaptive-TTA Mix 2* and *Adaptive-TTA Contrast*. Given the randomness inherent to some transformation parameters, some results are presented in the form of box plots, resulting from 10 different experiments.

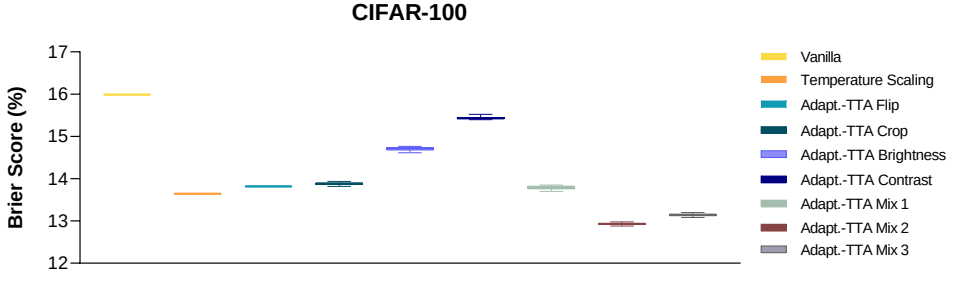

Figure 3: Results on the CIFAR-100 dataset. We refer to the caption of Figure 2 for detailed information about the experiments and results.

Figure 2 shows the results relative to the experiments made in the CIFAR-10 dataset, with respect to the effects of *Adaptive-TTA* (divided in seven different sub-methods, according to different augmentation policies used) in the uncertainty calibration of the DNN classifier, evaluated using the *Brier score*. It is noticeable that all presented approaches have better calibrated predictions than the *vanilla* procedure. *Adaptive-TTA Flip* and the *mixed* approaches (*Adaptive-TTA Mix 1*, *Mix 2* and *Mix 3)* achieved the best performance in terms of dealing with the predictive uncertainty of the model - with *Adaptive TTA Mix 1* being the best performing sub-method - all of them outperforming *temperature scaling*. From the remaining sub-methods, *Adaptive-TTA Crop* turns to be the most competitive of the three.

Figure 3 presents an analogous analysis of that shown in Figure 2, this time in the context of the CIFAR-100 dataset. Focusing primarily on the different sub-methods of *Adaptive-TTA*, we derive similar conclusions of those in the previous analysis, with *Adaptive-TTA Flip* and the *mixed* approaches having the strongest performance, but general good results across

the different augmentation policies deployed. Nonetheless, in this case, only *Adaptive-TTA Mix 2* and *Mix 3* outperform the *temperature scaling* baseline. Once again, *Adaptive-TTA Crop* also accomplishes competitive results.

5.2 AID

The AID [31] dataset comprises 10000 aerial scene RGB images with 600×600 pixels from Google Earth for the task of aerial scene classification, having 30 different classes. For the purpose of this work, this dataset is randomly divided into parcels of 70%, 10% and 20%, for training, validation and test sets, respectively. This dataset represents a fruitful addition to our analysis, since it introduces more real-world complexity and variability - when compared with the previous datasets - given its much greater resolution and the nature of its classes.

Figure 4 is analogous to the previous figures, this time presenting the results on the AID dataset. The first observation to be made is that the common pattern observed in the previously analyses is absent in this particular case. We verify that some of the approaches achieved results that are very close to the *vanilla* procedure, with *temperature scaling*, *Adaptive-TTA Flip* and *Adaptive-TTA Brightness* actually slightly aggravating the uncertainty calibration of the model (with *temperature scaling* having an increase in terms of *Brier score* of approximately 0.1%, being the worst performing method). We speculate that these differences in the results are caused by the aforementioned complexity and variability - present in this dataset - which probably caused some of the approaches to overfit their parameters to the characteristics of the validation set (even though in this case the validation set is proportionally bigger than those of the previous experiments). Nonetheless, it is possible to obtain satisfactory results with the proposed *mixed* approaches and, more importantly, *Adaptive-TTA Crop* accomplishes a relatively significant reduction of the uncertainty calibration of the DNN.

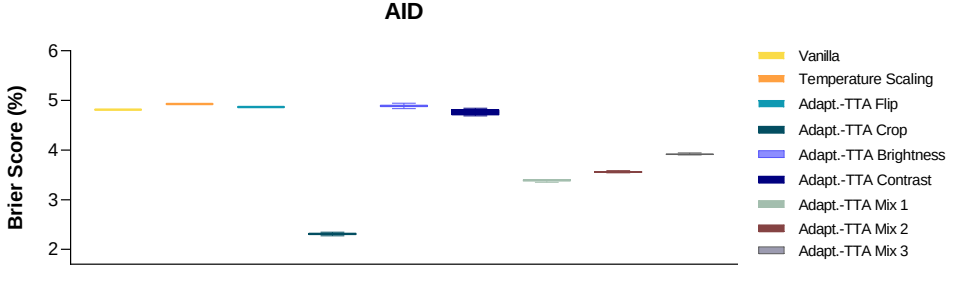

Figure 4: Results on the AID dataset. We refer to the caption of Figure 2 for detailed information about the experiments and results.

6 Final Remarks

In this work we introduced *Adaptive-TTA* - a novel approach to test time augmentation - that improves the uncertainty calibration of DNNs while preserving their accuracy. This method is evaluated with different augmentation policies, and is found to outperform *temperature*

scaling in the three datasets used in our experiments, with respect to the *Brier score*. Additionally, when analysing the results on the AID dataset, we verify that our method is capable of accomplishing good results in complex scenarios, where *temperature scaling* actually worsens the uncertainty calibration of the model.

Of the evaluated sub-methods, the *mixed* approaches were found to be the more consistent, with good results on all three datasets. Based on this evidence, we can conclude that is generally a good strategy to include different types of image transformations in the augmentation policy used in *Adaptive-TTA*. Also, the *crop* transformation is generally present in the best performing sub-methods, which introduces some evidence that this type of transformation preforms well with the proposed methodology.

Given the wide range of possibilities *Adaptive-TTA* can be applied (with multiple strategies in terms of augmentation policies), this work opens the possibility for even deeper research on the effects that different augmentation policies can have in this context. Furthermore, it would be interesting to find further insights on the reasons why different augmentation policies produce different results depending on the dataset they are applied, and hopefully introduce an unifying framework for the application of *Adaptive-TTA*. Also, the optimization process of the parameters of our method opens further possibilities for future experiments.

Acknowledgments

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

References

[1] Gail Blattenberger and Frank Lad. Separating the brier score into calibration and refinement components: A graphical exposition. *The American Statistician*, 39(1):26–32, 1985.

[2] Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[3] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[4] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[5] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330. PMLR, 2017.

[7] Chirag Gupta and Aaditya K Ramdas. Top-label calibration and multiclass-to-binary reductions. *arXiv preprint arXiv:2107.08353*, 2021.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[9] Sooyong Jang, Insup Lee, and James Weimer. Improving classifier confidence using lossy label-invariant transformations. In *International Conference on Artificial Intelligence and Statistics*, pages 4051–4059. PMLR, 2021.

[10] Ildoo Kim, Younghoon Kim, and Sungwoong Kim. Learning loss for test-time augmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:4163–4174, 2020.

[11] Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:18237–18248, 2020.

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[13] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[14] Alexander Lyzhov, Yuliya Molchanova, Arsenii Ashukha, Dmitry Molchanov, and Dmitry Vetrov. Greedy policy search: A simple baseline for learnable test-time augmentation. In *Conference on Uncertainty in Artificial Intelligence*, pages 1308–1317. PMLR, 2020.

[15] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning (ICML)*, pages 7034–7044. PMLR, 2020.

[16] Nikita Moshkov, Botond Mathe, Attila Kertesz-Farkas, Reka Hollandi, and Peter Horvath. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific reports*, 10(1):1–7, 2020.

[17] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.

[18] Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

[19] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[20] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[21] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 625–632, 2005.

[22] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

[23] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[24] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[25] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[26] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. When and why test-time augmentation works. *arXiv preprint arXiv:2011.11156*, 2020.

[27] Junjiao Tian, Dylan Yung, Yen-Chang Hsu, and Zsolt Kira. A geometric perspective towards neural calibration via sensitivity decomposition. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:26358–26369, 2021.

[28] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation. *Medical Imaging with Deep Learning (MIDL)*, 2018.

[29] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

[30] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[31] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55 (7):3965–3981, 2017.

[32] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and Naive Bayesian classifiers. In *International Conference on Machine Learning (ICML)*, volume 1, pages 609–616. Citeseer, 2001.

[33] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.