INSTITUTO DE SISTEMAS E ROBÓTICA
UNIVERSIDADE DE COIMBRA

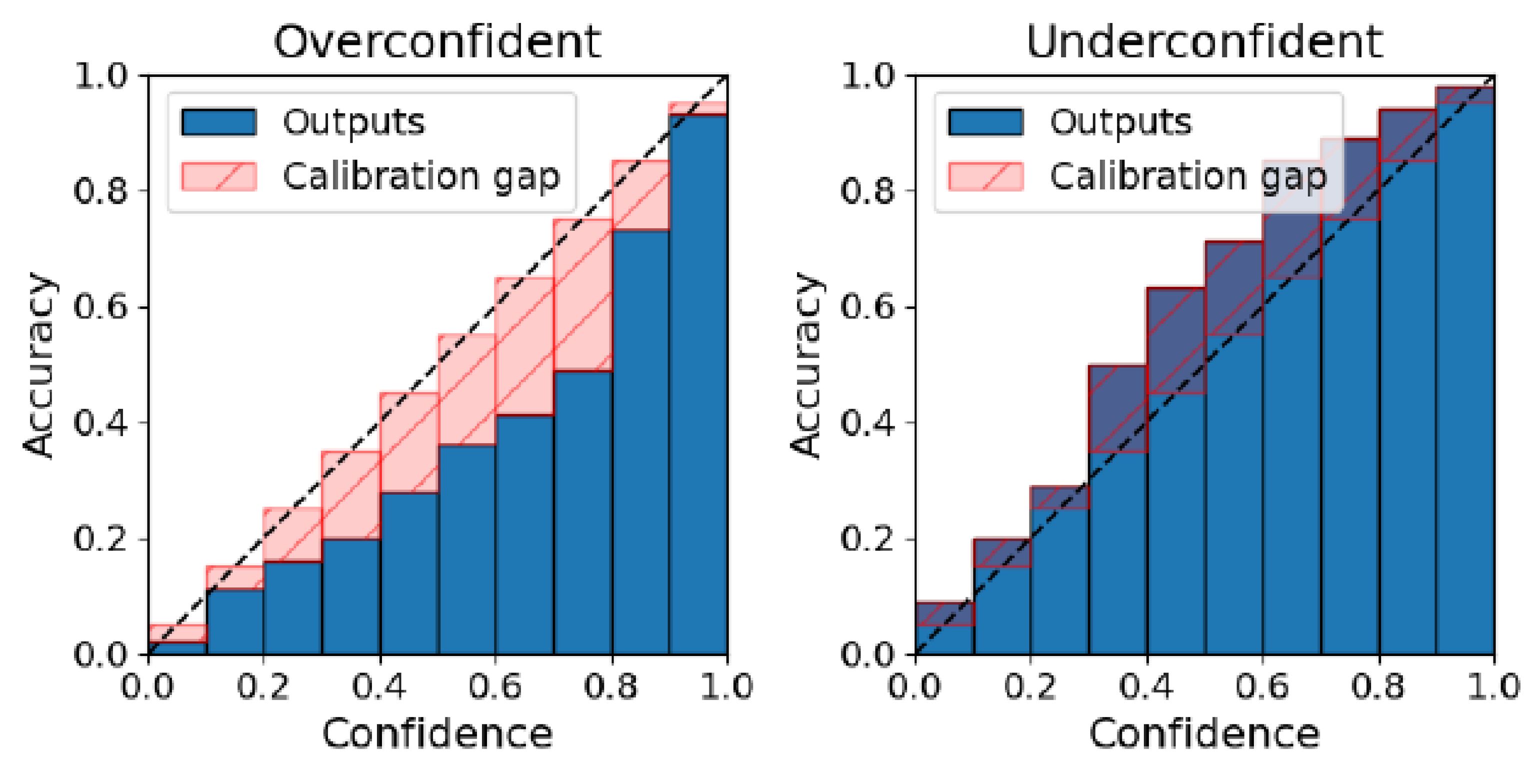
Improving the reliability and probabilistic interpretation of deep neural networks for critical applications

Pedro Conde

Supervisor: Cristiano Premebida

Critical applications of deep neural networks (DNNs) - like health, autonomous driving, disaster response and resource management - require a thorough look into the reliability of the proposed models. In addition to having highly accurate classification models, the user must be able to "trust" its predictions when dealing with critical scenarios.

Although much success as been achieved for a multitude of applications due to the accuracy of modern DNNs, this deep models have been found to be tendentially *uncalibrated*. This means that the confidence output generated by the model does not realistically represent the correct likelihood of its



***State-of-the-art*:**

Post-hoc Calibration

Recalibrating the outputs of the DNN without requiring any type of retraining. Usually highly dependent of a validation set.

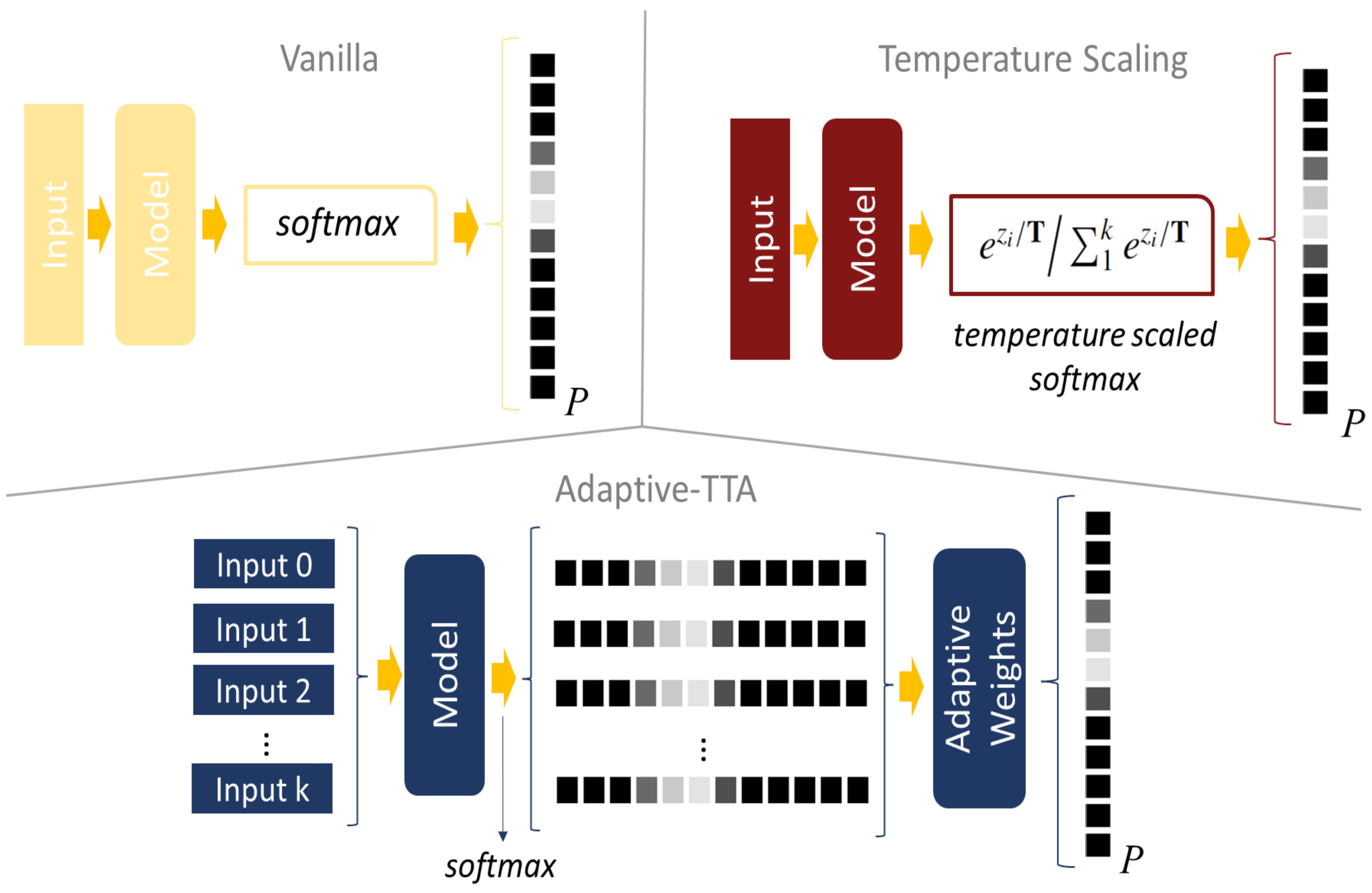
Probabilistic Models

Estimating model uncertainty using approximate bayesian inference. Computationally heavy when dealing with complex models.

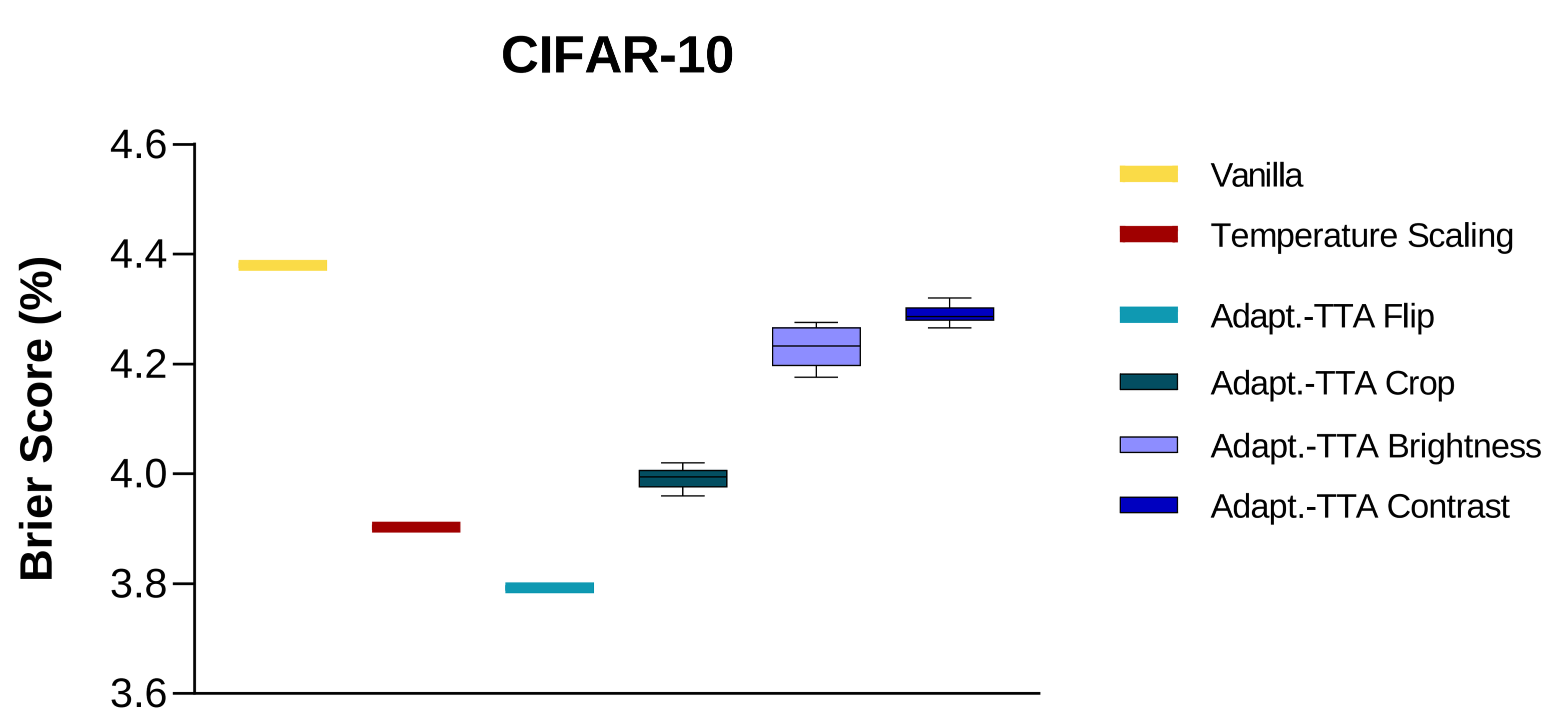
Confidence Estimation

Learning confidence estimates in parallel with the DNN's learning process. Still a recent field of research.

We are currently developing a technique called ***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.



Preliminary results already show that this technique is competitive with *state-of-the-art* post-hoc calibration methods, like *temperature scaling*, even surpassing their performance in some cases.



Acknowledgments:

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