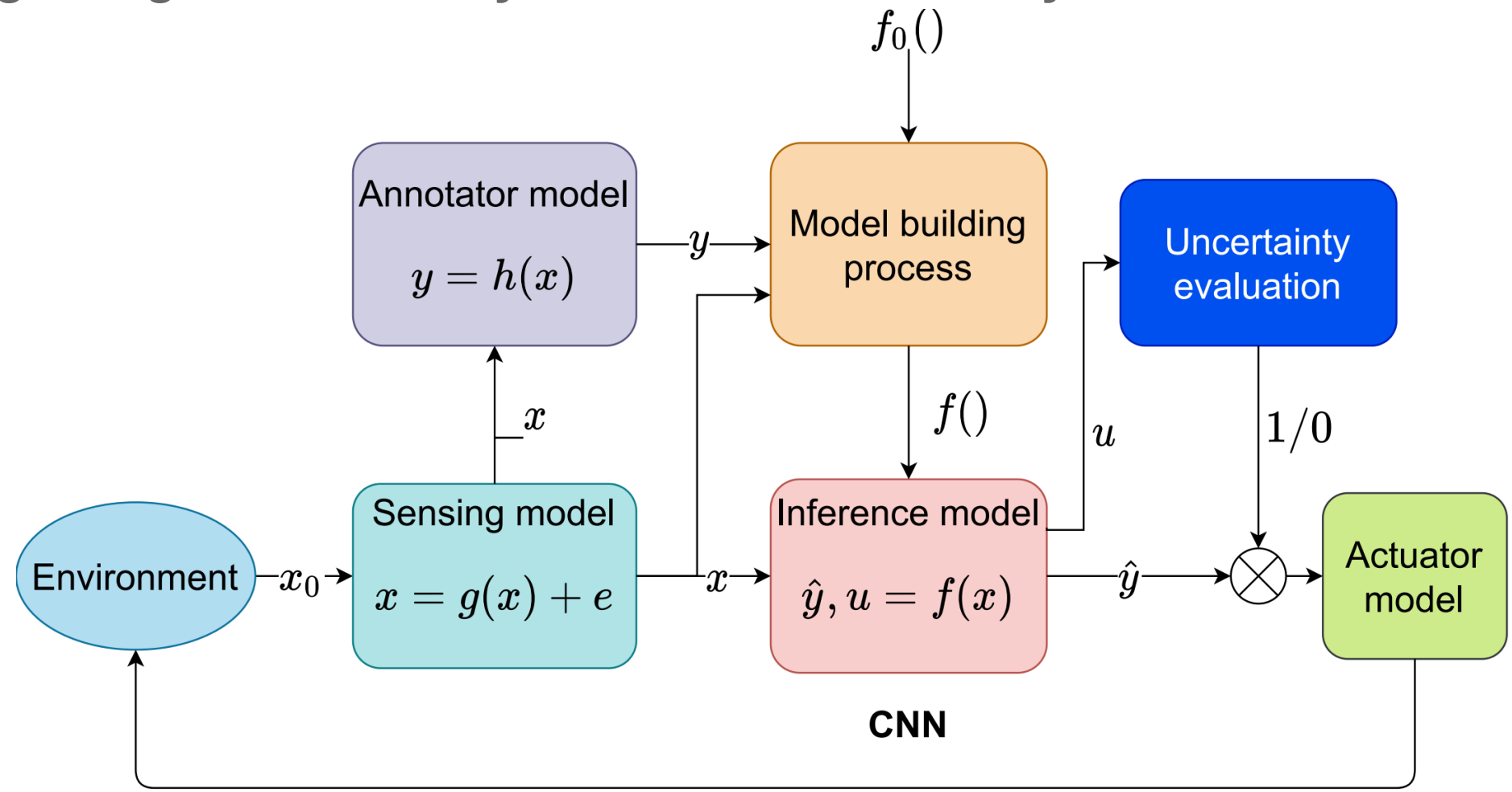
The impact of averaging logits over probabilities on ensembles of neural networks

Cedrique R. N. Tassi (Institute of Optical Sensor Systems)
Jakob Gawlikowski and **Auliya U. Fitri** (Institute of Data Science)
Rudolph Triebel (Institute of Robotics and Mechatronics)
German Aerospace Center (DLR)

Motivation: A convolutional neural network (CNN) classifier as part of an intelligent agent of a safety- or mission-critical system

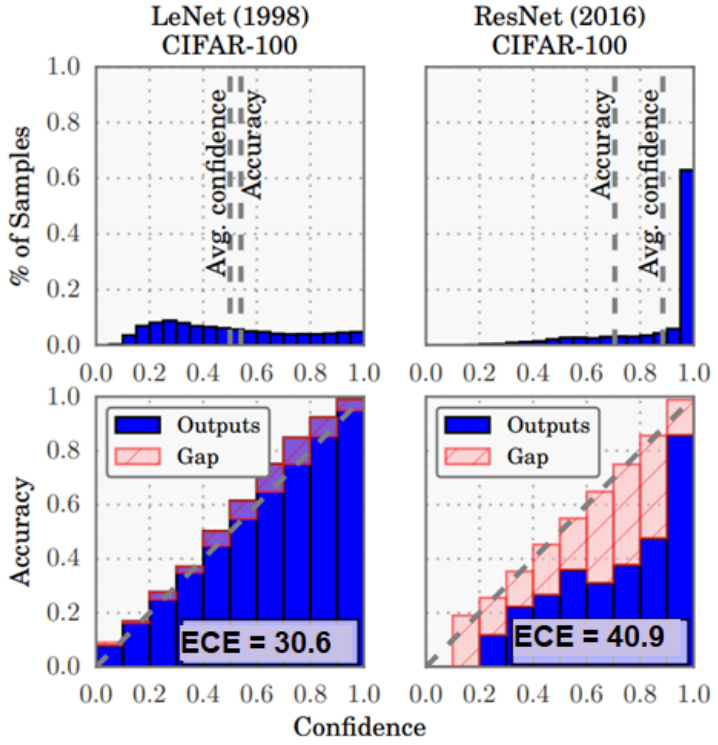


Problem formulation and solution approach

- **Given** a method for estimating uncertainty such as model averaging, the quality of uncertainty estimates is assessed by:
  - The ability to separate true predictions (TPs) and false predictions (FPs) → *We want high confidence (e.g. [5-1]) on TPs and low confidence (e.g. [0-5[) on FPs.*
  - The degree of calibration → *We want low calibration error on test data.*

- **However,** recent findings [Wen et al., 2020; Yao Qin et al., 2021; Rahaman and Thiery, 2021] show that model averaging does not necessarily lead to calibrated confidences.

- While existing methods for improving calibration in model averaging focus on recalibrating [Rahaman and Thiery, 2021], building [Wen et al., 2020] , or sampling [Zhang et al., 2019] calibrated models, **we rethink the combination process in model averaging**.



[Guo et al., 2017]

Definition and related works

The combination process describes how the predictions of multiple members are combined and the information type (e.g., logits or probabilities) that is combined.

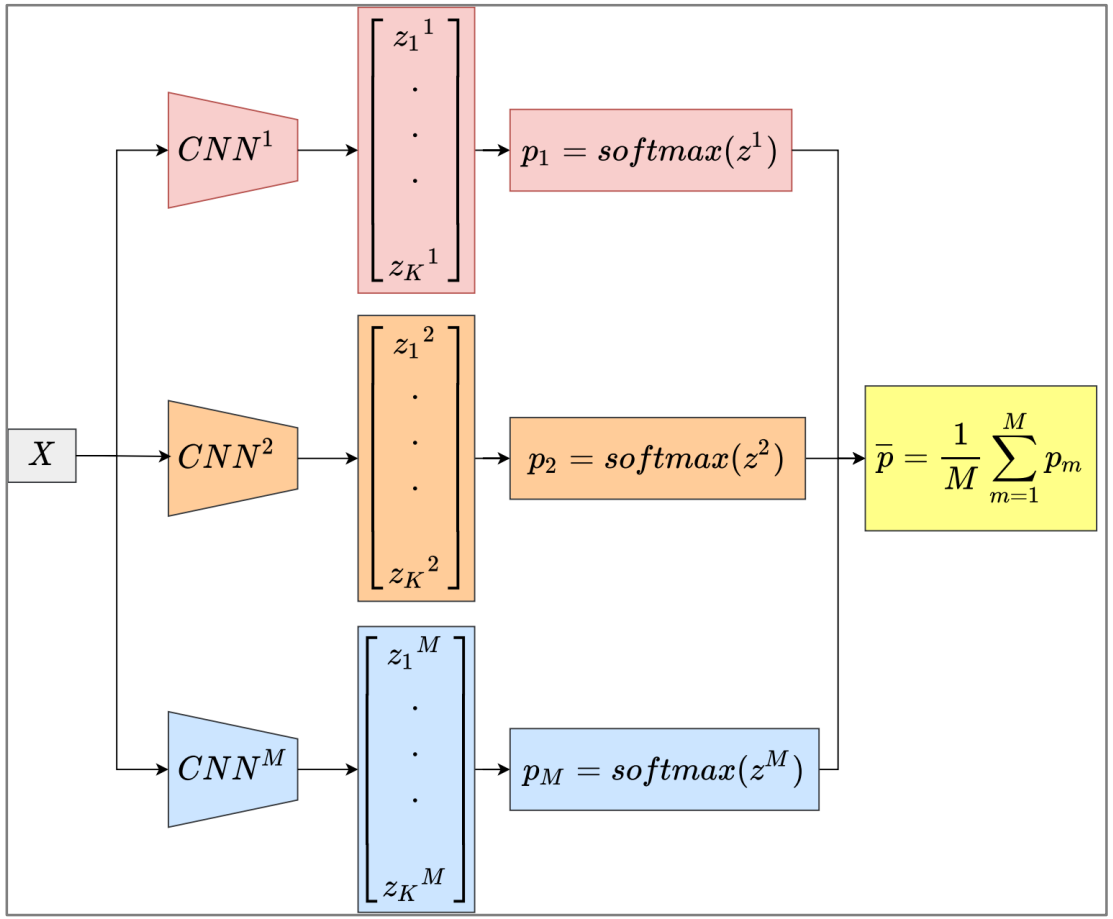
Works studying combination approaches

- Stacking [Ju et al., 2018]
- Plurality, majority and other voting methods [Kuncheva, 2014; Van Erp et al., 2002; Tajti, 2020]
- [Van Erp et al., 2002; Ju et al., 2018] showed that whether one approach improves accuracy better than another depends on several factors, such as the number of members, the diversity inherent in individual members, and the accuracy of individual members.
- [Tassi, 2019] showed that the averaging approach captures uncertainty better than voting approaches such as plurality or majority voting. This supports the finding of [Clemen,1989] and [Kittler et al., 1998].

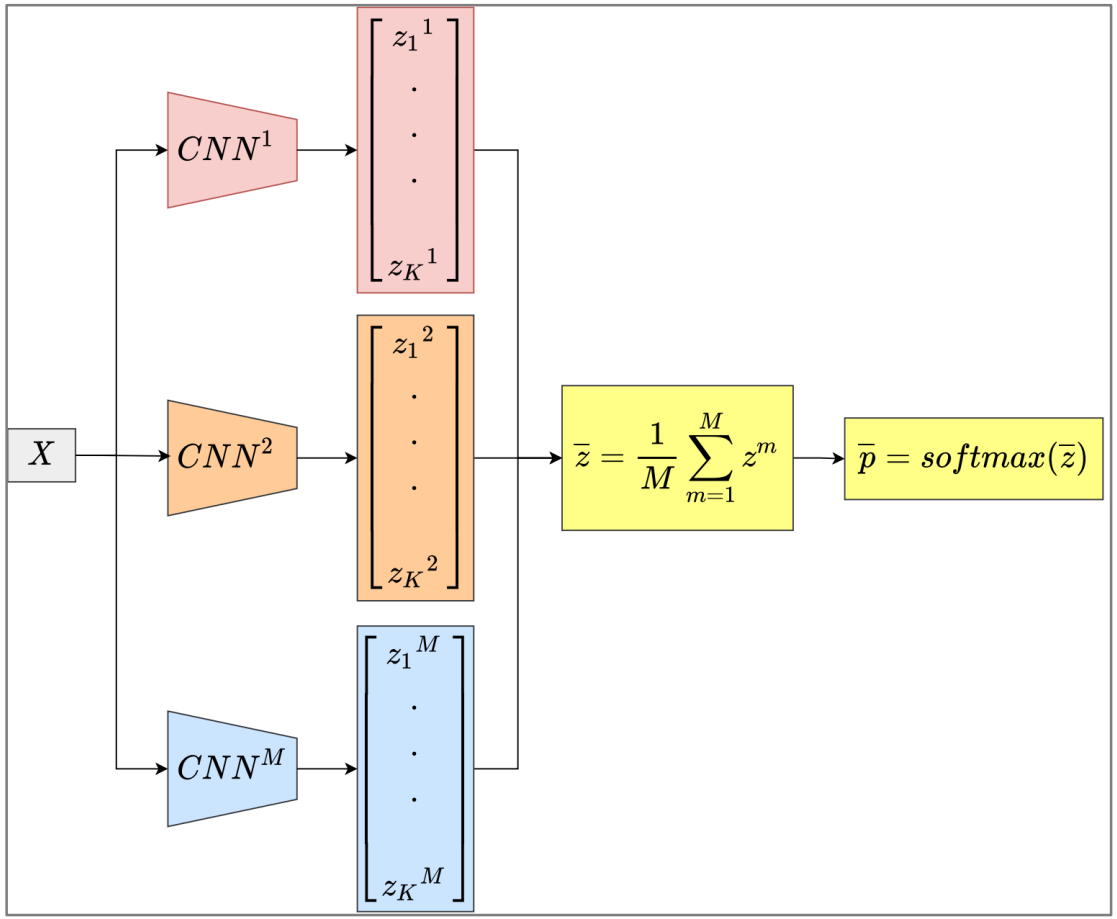
Works studying the information type

- [Ju et al., 2018] evaluated the impact of combining logits instead of probabilities on accuracy, but the impact on the quality of confidence remains unclear.
- *We evaluated the impact of logit-averaging on the quality of confidence.*

Averaging logits instead of probabilities in an ensemble

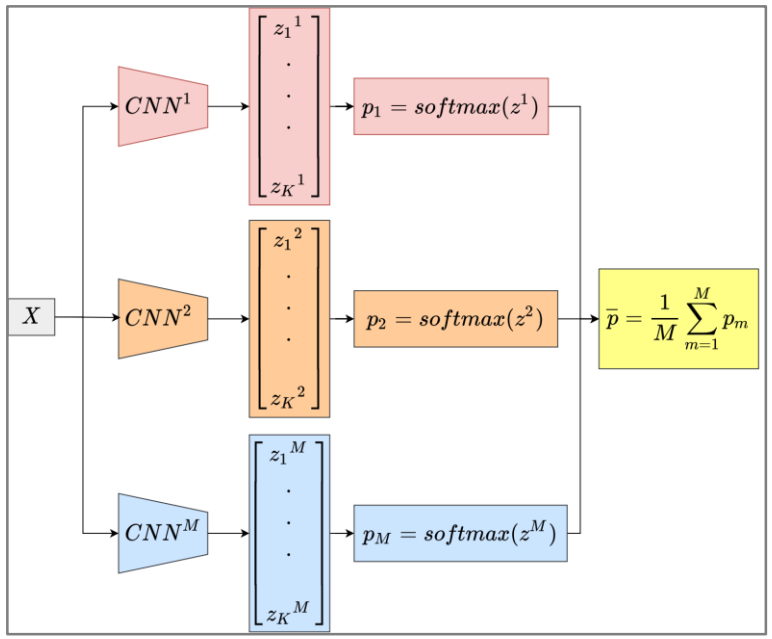


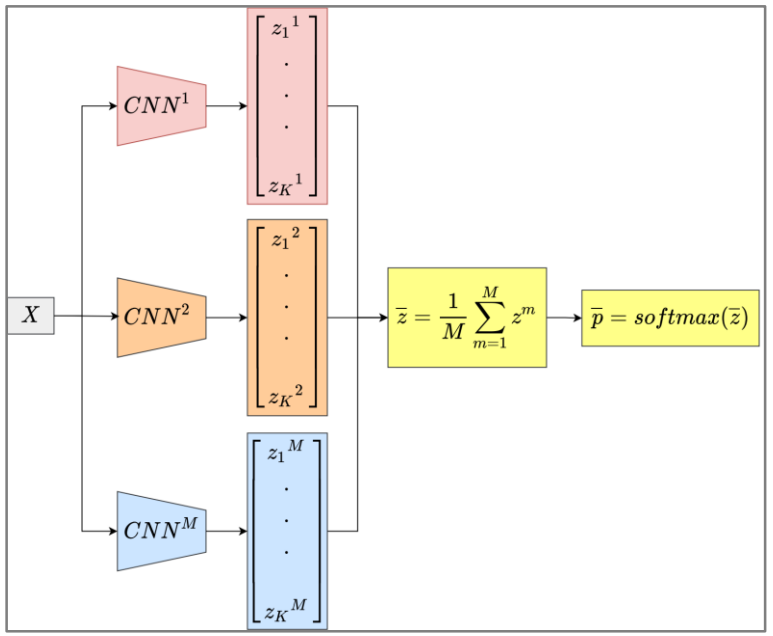
Probability averaging



Logit averaging

Compare logit-averaging and probability-averaging





Probability averaging	Logit averaging
Inductive biases limit the confidence	Reduce inductive biases and therefore increases the confidence
Gives the best confidence associated with decisions made based on weak evidence (biased logits)	Gives the best evidence (unbiased logit) for decision making
Suite for decision making **under uncertainty**	Suite for decision making **under no uncertainty**

Experiments: The impact of averaging logits instead of probabilities on accuracy and calibration error

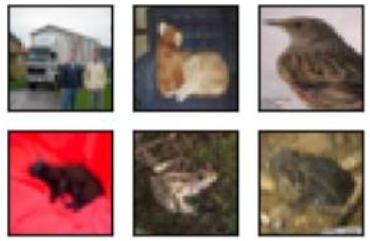
Comparison of accuracy[%], average confidence[%] (in bracket), and expected calibration error (ECE) $[10^{-2}]$ of ensembles using two approaches for averaging underconfident CNNs trained using strong regularization: average probabilities (AP) and average logits (AL). The results were obtained using the test data of experimental data-sets.

	Accuracy (Average confidence) ↑		**ECE** ↓	
	AP	AL	AP	AL
CIFAR10 (DenseNets)	89.52 (84.31)	89.60 (**87.97**)	5.23	**2.47**
FashionMNIST (ResNets)	92.70 (87.86)	92.58 (**90.16**)	5.15	**2.86**
MNIST (VGGNets)	99.04 (98.24)	99.04 (**98.89**)	1.03	**0.52**

Comparison of accuracy[%], average confidence[%] (in bracket), and expected calibration error (ECE) $[10^{-2}]$ of ensembles using two approaches for averaging overconfident CNNs trained using weak regularization: average probabilities (AP) and average logits (AL). The results were obtained using the test data.

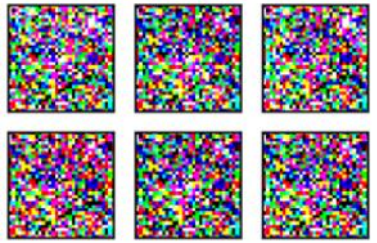
	Accuracy (Average confidence) ↑		**ECE** ↓	
	AP	AL	AP	AL
CIFAR10 (DenseNets)	88.67 (89.43)	88.88 (**96.17**)	**3.03**	7.40
FashionMNIST (ResNets)	94.49 (95.86)	94.58 (**98.43**)	**1.98**	4.11

Experiments: The impact of averaging logits instead of probabilities on the ability to separate TPs and FPs



(a) Test data

(b) Swapped data



(c) Noisy data

Comparison of average confidence[%] of ensembles using two approaches (average probabilities (AP) and average logits (AL)) for averaging underconfident networks trained using strong regularization and evaluated on TPs and FPs: TPs were obtained on *subsets of the correctly classified test data*, while FPs were obtained on *swapped*, *noisy* and out-of-domain (OOD) data.

	TP ↑		**FP (OOD) ↓**		**FP (Swapped) ↓**		**FP (Noisy) ↓**	
	AP	AL	AP	AL	AP	AL	AP	AL
CIFAR10 (DenseNets)	93.94	**96.63**	**35.39**	40.08	**51.84**	56.03	**39.42**	58.69
FashionMNIST (ResNets)	88.01	**90.16**	**55.48**	63.21	**59.30**	67.91	**81.39**	99.82
MNIST (VGGNets)	99.09	**99.55**	**57.16**	80.45	**51.96**	62.01	**69.58**	88.84

Summary and implications

- We showed that logit-averaging preserves accuracy but increases the degree of confidence on both TPs and FPs.

- We also showed that the increase in the degree of confidence caused by logit-averaging reduces the calibration error for underconfident networks but increases it for overconfident networks. → *This suggests that we can apply logit-averaging in an ensemble of underconfident networks to reduce the calibration error.*

- However, we showed that the increase in the degree of confidence caused by logit-averaging can harm the separability between TPs and FPs. → *This finding suggests that for safety- and mission-critical applications where the separability of TPs and FPs is essential, we should apply probability-averaging as it is usually done.*

Thanks for your attention!

Any questions ?

References

[Clemen,1989]	Clemen, Robert T. "Combining forecasts: A review and annotated bibliography." International journal of forecasting 5.4 (1989): 559-583.
[Kittler et al., 1998]	Kittler, Josef, et al. "On combining classifiers." IEEE transactions on pattern analysis and machine intelligence 20.3 (1998): 226-239.
[Van Erp et al., 2002]	Van Erp, Merijn, Louis Vuurpijl, and Lambert Schomaker. "An overview and comparison of voting methods for pattern recognition." Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition. IEEE, 2002.
[Kuncheva, 2014]	Kuncheva, Ludmila I. Combining pattern classifiers: methods and algorithms. John Wiley & Sons, 2014.
[Tajti, 2020]	Tajti, Tibor. "New voting functions for neural network algorithms." Annales Mathematicae et Informaticae. Vol. 52. Eszterházy Károly Egyetem Líceum Kiadó, 2020.
[Ju et al., 2018]	Ju, Cheng, Aurélien Bibaut, and Mark van der Laan. "The relative performance of ensemble methods with deep convolutional neural networks for image classification." Journal of Applied Statistics 45.15 (2018): 2800-2818.
[Wen et al., 2020]	Wen, Yeming, et al. "Combining ensembles and data augmentation can harm your calibration." arXiv preprint arXiv:2010.09875 (2020).
[Yao Qin et al., 2021]	Qin, Yao, et al. "Improving calibration through the relationship with adversarial robustness." Advances in Neural Information Processing Systems 34 (2021): 14358-14369.
[Rahaman and Thiery, 2021]	Rahaman, Rahul. "Uncertainty quantification and deep ensembles." Advances in Neural Information Processing Systems 34 (2021): 20063-20075.
[Zhang et al., 2019]	Zhang, Zhilu, Adrian V. Dalca, and Mert R. Sabuncu. "Confidence calibration for convolutional neural networks using structured dropout." arXiv preprint arXiv:1906.09551 (2019).
[Guo et al., 2017]	Guo, Chuan, et al. "On calibration of modern neural networks." International conference on machine learning. PMLR, 2017.
[Tassi, 2019]	Tassi, Njieutcheu, and Cedrique Rovile. "Bayesian convolutional neural network: Robustly quantify uncertainty for misclassifications detection." Mediterranean Conference on Pattern Recognition and Artificial Intelligence. Springer, Cham, 2019.