

SUPPLEMENTARY MATERIAL FOR IMPROVING CALIBRATION THROUGH THE RELATIONSHIP WITH ADVERSARIAL ROBUSTNESS

Anonymous authors

Paper under double-blind review

A IMPLEMENTATION DETAILS

A.1 SVHN

We use a simple network architecture with 7 convolutional layers and follow all the training details introduced in Qin et al. (2020) for SVHN. This network architecture achieves the state-of-the-art accuracy on SVHN.

A.2 CIFAR-10

All the experimental results on CIFAR-10 were obtained with a ResNet-29 v2 (He et al., 2016b) with a batch size of 256. The network is trained with Adam optimizer (Kingma et al., 2015) for 200 epochs. The initial learning rate is 10^{-3} and decayed down to 10^{-4} after 80 epochs, 10^{-5} after 120 epochs, 10^{-6} after 160 epochs and 0.5×10^{-6} after 180 epochs. We adapted the following data augmentation and training script at https://keras.io/examples/cifar10_resnet/. The training mechanism is the same for all the methods that we compare in the main paper. We randomly split the training dataset into training data of 45000 images and 5000 images as the validation set. The test set has 10000 images.

For label smoothing (LS), we sweep the hyperparameter ϵ within the range $[0, 0.1]$ with a step size 0.01 and find that the network has the best calibration performance on the validation set when $\epsilon = 0.02$.

For Adaptive Label Smoothing (AdaLS), there is a hyperparameter α which plays a role as learning rate in the adaptive learning mechanism. We choose hyperparameter α based on the calibration performance on the validation set. Specifically, we run experiments with $\alpha \in \{0.005, 0.01, 0.05, 0.1\}$ and find that $\alpha = 0.05$ achieve the best calibration performance.

Similarly, for Adversarial Robustness based Adaptive Label Smoothing (AR-AdaLS), we choose the hyperparameter α from the set $\{0.005, 0.01, 0.05\}$ and empirically set $\alpha = 0.005$ which has the best calibration performance on the validation set. We use the same hyperparameter $\alpha = 0.005$ without further tuning for AR-AdaLS of Ensemble.

All the results of ensemble models are obtained via training 5 independent models with random initializations.

A.3 CIFAR-100

We train a Wide ResNet-28-10 v2 (Zagoruyko & Komodakis, 2016) to obtain the state-of-the-art accuracy for CIFAR-100. We adapt the same training details and data augmentation at <https://github.com/google/edward2/blob/master/baselines/cifar/deterministic.py>.

For label smoothing, we sweep the hyperparameter ϵ within the range $[0, 0.1]$ with a step size 0.01 and find that the network has the best calibration performance on the validation set when $\epsilon = 0.07$.

The hyperparameter α is set to be 0.005 in AR-AdaLS. For AR-AdaLS that generated with on-the-fly adversarial examples, we recompute the adversarial robustness for training and validation sets after 65, 130 epochs.

A.4 IMAGENET

All the experiments on ImageNet were obtained via training a ResNet-101 v1 (He et al., 2016a) following the training script at <https://github.com/google/edward2/blob/master/baselines/imagenet/deterministic.py>. The network is trained with a batch size of 128 for each TPU core with SGD optimizer for 90 epochs. The input image is normalized (divided by 255) to be within $[0,1]$. We randomly divide 50000 validation images into validation set with 25000 images and test set with 25000 images. Note that the same dataset and training mechanisms are used for all the methods that we compare in the main paper.

For Label Smoothing (LS), we sweep the hyperparameter ϵ within the range $[0, 0.1]$ with a step size 0.01 and find that the best calibration performance on the validation set is achieved by setting $\epsilon = 0.02$.

For Adaptive Label Smoothing (AdaLS), we sweep the hyperparameter α in the set $\{0.005, 0.01, 0.03, 0.05, 0.1\}$ and set it to be $\alpha = 0.03$ for the best calibration performance on the validation set.

We empirically set $\alpha = 0.001$ for AR-AdaLS in the first 60 epochs of the training and then increase it to 0.05 for the next 30 epochs. The same hyperparameter α is used for AR-AdaLS of Ensemble without further tuning.

All the ensemble models are a combination of 5 independent models with random initializations.

A.5 CW ATTACKS

To compute the adversarial robustness, we construct ℓ_2 based CW attacks (Carlini & Wagner, 2017) following the code at https://github.com/tensorflow/cleverhans/blob/master/cleverhans/attacks/carlini_wagner_l2.py. Specifically, we set the binary search steps to be 3, max iterations to be 500 and learning rate to be 0.005. The generated untargeted CW attacks can achieve 100% success rate for all the datasets that we consider: SVHN, CIFAR-10, CIFAR-100 and ImageNet. We set the number of adversarial robustness training subset and validation subset to be $R = 10$ respectively.

B CALIBRATION PERFORMANCE: ADDITIONAL RESULTS

In Figure 1, we show ECE and accuracy of all the single-models and their corresponding ensembles on the clean test and shifted CIFAR-10 and ImageNet. The high-level conclusion is: 1) all the ensemble models that we compare have similar accuracies, which are higher than single-models. 2) all the ensemble methods perform relatively well for highly shifted data but ensembles of well calibrated models (Ensemble of LS, Ensemble of AdaLS, Ensemble of AR-AdaLS) perform much worse on less shifted and clean test data. 3) AR-AdaLS of Ensemble successfully learns to adaptively adjust its smoothing level to keep the ensemble well calibrated on both clean and shifted dataset.

Figure 3 is to validate the observation that naive ensembling of a single model can improve calibration on highly shifted data where single-model is over-confident but can harm calibration on clean test data where single model is well calibrated. As shown in Figure 3, we can see that if the single-model is over-confident (Vanilla), Ensemble of Vanilla greatly improves the model’s calibration. However, when the single model is well-calibrated (LS, AdaLS, AR-AdaLS), naive ensembling of these models leads to an under-confident model with a worse calibration performance. In contrast, AR-AdaLS of Ensemble successfully avoids this issue by adaptively smoothing the training labels according to adversarial robustness of training data.

C NUMERICAL RESULTS

We report the mean of variance and mean of ECE across 19 types of shift for CIFAR-10-C and 15 types of shift for ImageNet-C in Table 1 and Table 2. We can see that AR-AdaLS achieves the best performance in model’s calibration (the smallest ECE) and stability (the smallest variance) among all the single-model based methods on the corrupted datasets. For the ensemble models, our proposed AR-AdaLS of Ensemble also achieves the best calibration performance on the corrupted datasets.

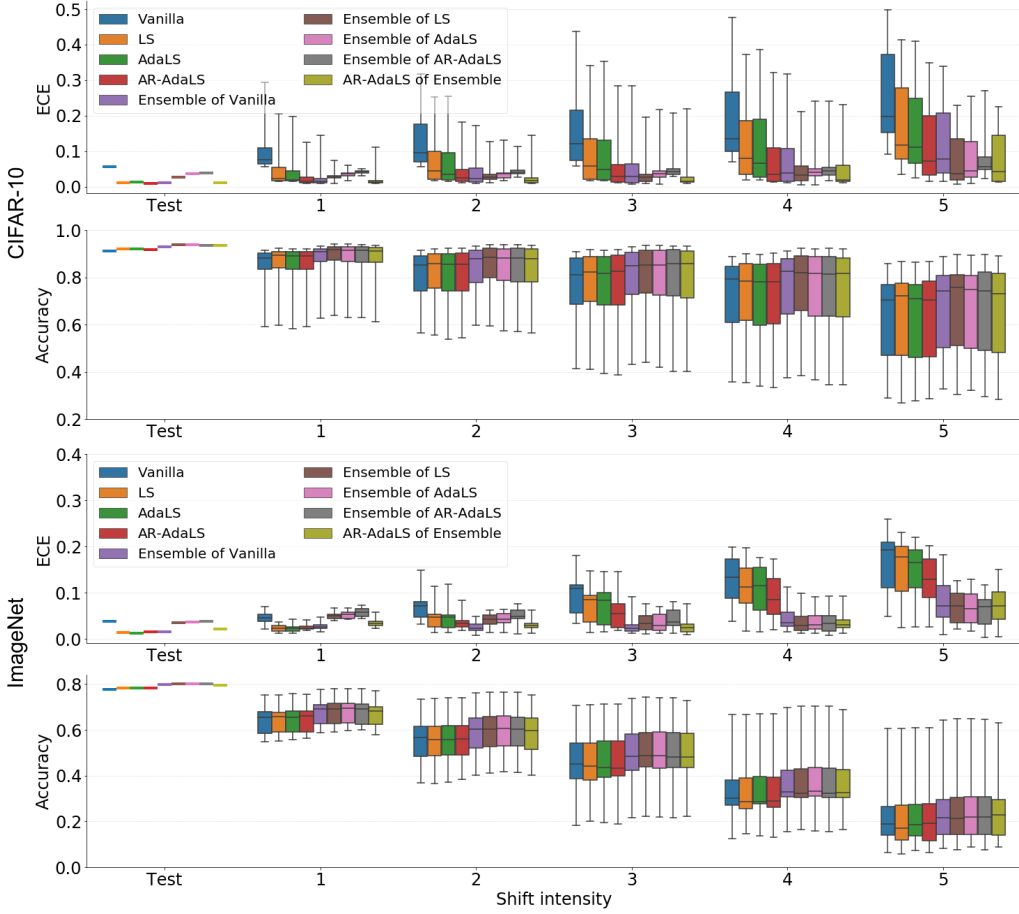


Figure 1: Comparison of ensembled models: ECE and Accuracy on both clean test data and shifted data on CIFAR-10 and ImageNet. For each intensity of shift, we show the results with a box plot summarizing the 25th, 50th, 75th quartiles across 19 types of shift on CIFAR-10-C and 15 types of shift on ImageNet-C. The error bars indicate the *min* and *max* value across different shift types.

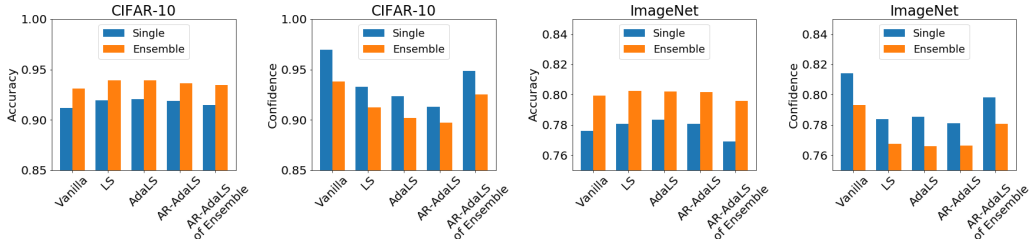


Figure 2: Comparing accuracy and confidence of the predicted class between single model and the corresponding ensemble model for each method.

REFERENCES

- Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.
- 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*, pp. 770–778, 2016a.

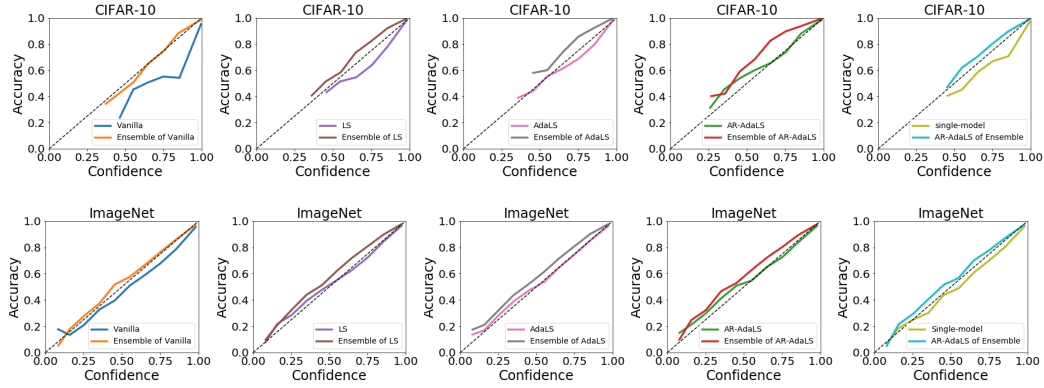


Figure 3: Reliability diagram of accuracy versus confidence of single model and ensemble model on the clean test of CIFAR-10 and ImageNet. The perfect calibrated model should be aligned with the diagonal dotted line (above is under-confident, below is over-confident).

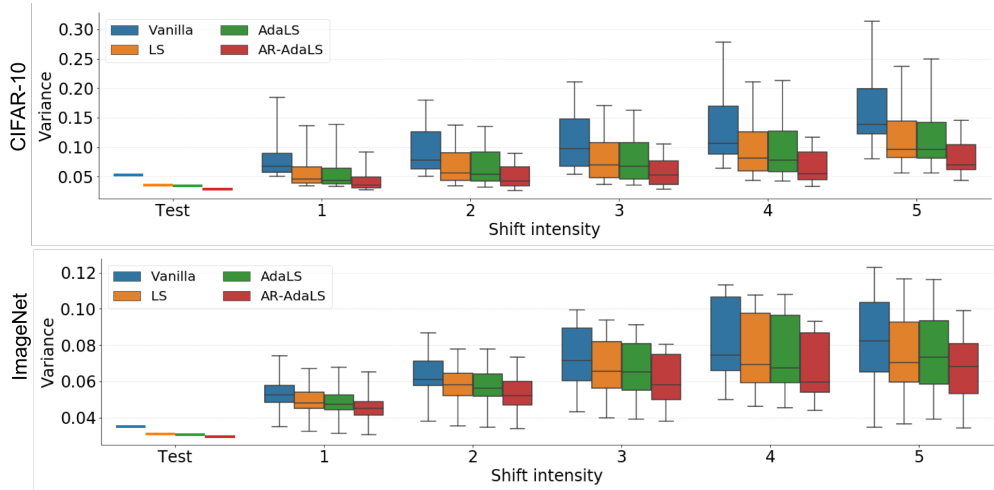


Figure 4: Variance on clean test and shifted data on CIFAR-10 and ImageNet. For each shift intensity, we show the results with a box plot summarizing the 25th, 50th, 75th quartiles across 19 shift types on CIFAR10-C and 15 shift types on ImageNet-C. The error bars indicate the *min* and *max* value across different shift types.

Table 1: Mean of variance ($\times 10^{-2}$) across 19 types of shift for CIFAR-10-C and 15 types of shift for ImageNet-C. The best model is shown in **bold**.

Dataset	CIFAR-10-C					ImageNet-C				
Shift Intensity	1	2	3	4	5	1	2	3	4	5
Vanilla	7.85	9.69	11.2	13.1	16.0	5.28	6.39	7.37	8.23	8.29
LS	5.54	6.95	8.11	9.65	11.8	4.86	5.84	6.78	7.55	7.41
AdaLS	5.47	6.87	7.95	9.44	11.5	4.79	5.77	6.66	7.51	7.56
AR-AdaLS	4.21	5.06	5.73	6.66	8.24	4.53	5.49	6.12	6.76	6.66

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645. Springer, 2016b.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pp. 2575–2583, 2015.

Yao Qin, Nicholas Frosst, Colin Raffel, Garrison Cottrell, and Geoffrey Hinton. Deflecting adversarial attacks. *arXiv preprint arXiv:2002.07405*, 2020.

Table 2: Mean of ECE ($\times 10^{-2}$) across 19 types of shift for CIFAR-10-C and 15 types of shift for ImageNet-C. The best single model and ensemble model are shown in **bold** respectively.

Dataset	CIFAR-10-C					ImageNet-C				
Shift Intensity	1	2	3	4	5	1	2	3	4	5
Vanilla	9.59	12.9	15.6	19.5	25.7	4.67	6.94	9.60	13.2	16.6
LS	4.45	7.06	9.15	12.1	17.5	2.38	4.62	7.39	11.4	15.1
AdaLS	4.06	6.65	8.85	11.7	16.5	2.37	4.48	7.18	11.1	14.6
AR-AdaLS	2.57	4.09	5.64	7.85	11.9	2.44	3.58	5.80	9.33	12.9
Ensemble of Vanilla	2.47	4.13	5.73	8.03	12.1	2.77	2.51	2.87	4.64	8.26
Ensemble of LS	2.77	3.24	3.96	5.20	7.85	4.96	4.18	3.65	3.65	6.94
Ensemble of AdaLS	3.48	3.74	4.86	5.95	7.74	5.20	4.36	3.61	3.74	6.92
Ensemble of AR-AdaLS	4.23	4.52	5.23	5.68	7.68	5.76	5.23	4.51	3.69	6.33
AR-AdaLS of Ensemble	1.97	2.98	3.91	5.38	7.98	3.41	3.03	2.84	3.61	7.03

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *ArXiv*, abs/1605.07146, 2016.