Improving Calibration through the Relationship with Adversarial Robustness

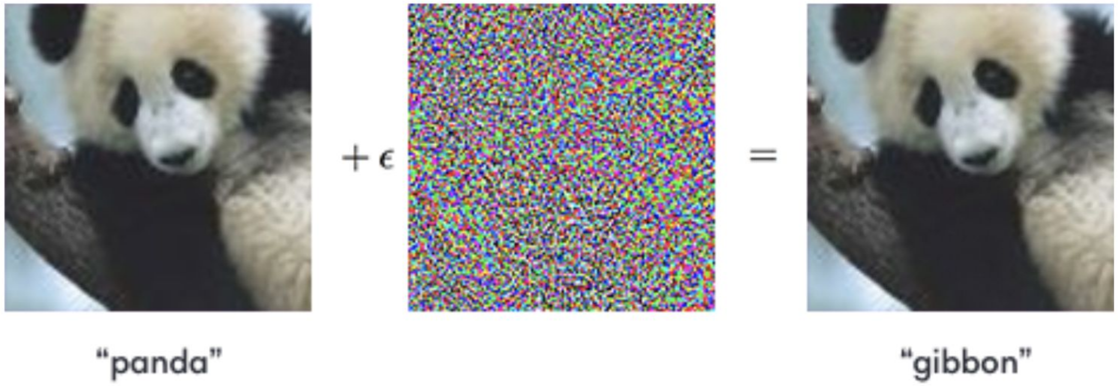
Yao Qin, Xuezhi Wang, Alex Beutel, Ed H. Chi

Google Research

Presenter: Yao Qin

What is Robustness?

- **Adversarial Robustness**
  - Neural networks lack *adversarial robustness*, i.e., small perturbations to inputs cause incorrect predictions.



What is Robustness?

- **Adversarial Robustness**
  - Neural networks lack *adversarial robustness*, i.e., small perturbations to inputs cause incorrect predictions.
- **Calibration**
  - Neural networks are often *miscalibrated*, i.e., the predicted probability is not a good indicator of how much we should trust our model.

Google Research

What is Robustness?

- **Adversarial Robustness**
  - Neural networks lack *adversarial robustness*, i.e., small perturbations to inputs cause incorrect predictions.
- **Calibration**
  - Neural networks are often *miscalibrated*, i.e., the predicted probability is not a good indicator of how much we should trust our model.

Know what they do not know

What is Robustness?

- **Adversarial Robustness**
  - Neural networks lack *adversarial robustness*, i.e., small perturbations to inputs cause incorrect predictions.
- **Calibration**
  - Neural networks are often *miscalibrated*, i.e., the predicted probability is not a good indicator of how much we should trust our model.

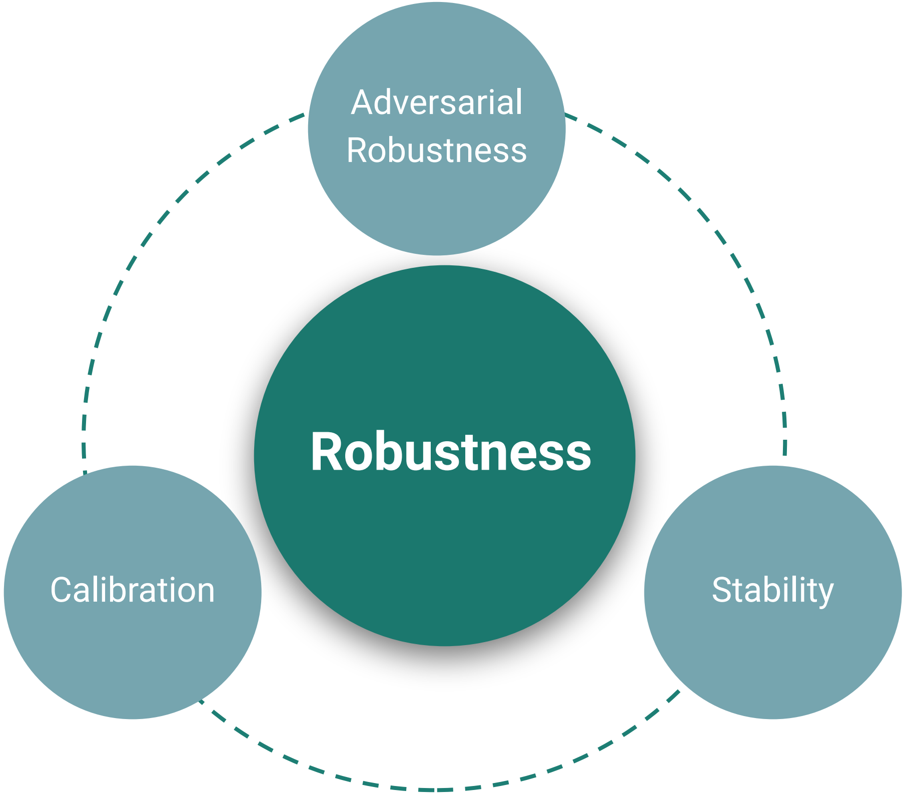
Know what they do not know

Over-confident! 🙁

What is Robustness?

- **Adversarial Robustness**
  - Neural networks lack *adversarial robustness*, i.e., small perturbations to inputs cause incorrect predictions.
- **Calibration**
  - Neural networks are often *miscalibrated*, i.e., the predicted probability is not a good indicator of how much we should trust our model.
- **Stability**
  - Neural networks give *unstable* predictions, i.e., the predicted probabilities vary greatly over multiple independent runs.

Any relationship between different “robustness”?



Quantify robustness

- **Adversarial Robustness**
  - Given an input x and a classifier $f(\cdot)$, we construct ℓ_2 norm based CW adversarial attack [1] that $f(x+\delta) \neq f(x)$.

Adversarial robustness = $||$Adversarial perturbation $\delta\ ||_2$

[1] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*

Google Research

Quantify robustness

- **Adversarial Robustness**
  - Given an input x and a classifier $f(\cdot)$, we construct ℓ_2 norm based CW adversarial attack [1] that $f(x+\delta) \neq f(x)$.

Adversarial robustness = $||$Adversarial perturbation δ $||_2$

$\boxed{\text{Larger Adv. perturbation} \Longrightarrow \text{More Adv. robust input } x}$

[1] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*

Google Research

Quantify robustness

- **Adversarial Robustness** (Larger Adv. perturbation $\Longrightarrow$ More Adv. robust input x)
- **Calibration**
  - Expected calibration error (ECE) measures how well accuracy and confidence of the predicted class are aligned [1].

Calibration Error = |Confidence - Accuracy|

predicted probability of correctness

observed frequency of correctness

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. *ICML* 2017.

Quantify robustness

- **Adversarial Robustness** (Larger Adv. perturbation ⇒ More Adv. robust input x)
- **Calibration**
  - Expected calibration error (ECE) measures how well accuracy and confidence of the predicted class are aligned [1].

Calibration Error = |Confidence - Accuracy|

predicted probability of correctness

observed frequency of correctness

Larger ECE ⇒ Worse calibrated prediction $f(x)$

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. *ICML* 2017.

Google Research

Quantify robustness

- **Adversarial Robustness** (Larger Adv. perturbation $\Rightarrow$ More Adv. robust input x)
- **Calibration** (Larger ECE $\Rightarrow$ Worse calibrated prediction $f(x)$)
- **Stability**
  - Variance of the predicted probability of multiple independent runs with random initialization [1].

Larger variance $\Rightarrow$ Less stable prediction $f(x)$

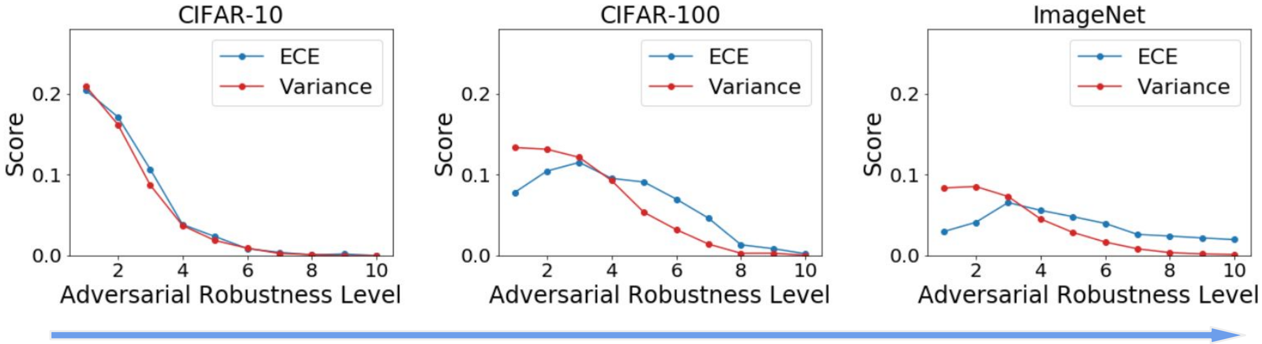
[1] T. Pearce, A. Brintrup, M. Zaki, and A. Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. *ICML* 2018.

Correlation

- **Adversarial Robustness** (Larger Adv. perturbation $\Rightarrow$ More Adv. robust input x)
- **Calibration** (Larger ECE $\Rightarrow$ Worse calibrated prediction $f(x)$)
- **Stability** (Larger variance $\Rightarrow$ Less stable prediction $f(x)$)

Correlation

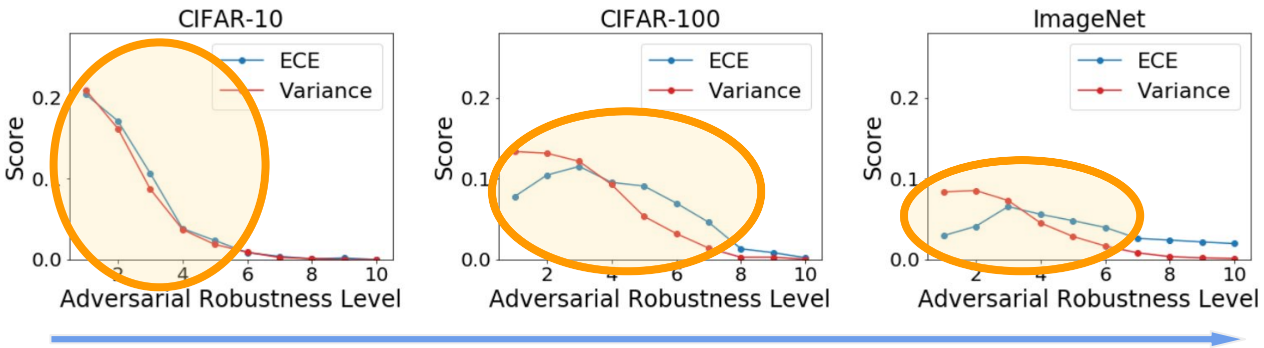
- **Adversarial Robustness** (Larger Adv. perturbation $\Rightarrow$ More Adv. robust input x)
- **Calibration** (Larger ECE $\Rightarrow$ Worse calibrated prediction $f(x)$)
- **Stability** (Larger variance $\Rightarrow$ Less stable prediction $f(x)$)



Larger adversarial robustness level → More adv. robust input

Correlation

- **Adversarial Robustness** (Larger Adv. perturbation ⇒ More Adv. robust input x)
- **Calibration** (Larger ECE ⇒ Worse calibrated prediction $f(x)$)
- **Stability** (Larger variance ⇒ Less stable prediction $f(x)$)



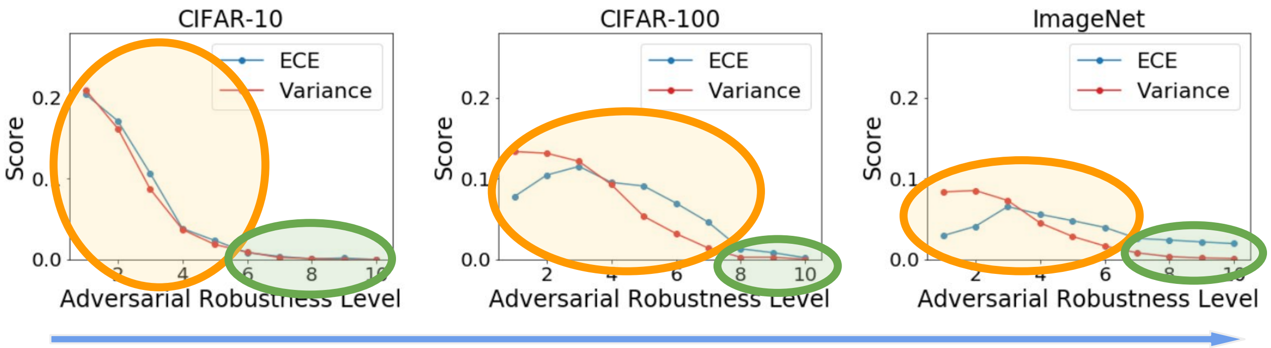
Larger adversarial robustness level → More adv. robust input

Less adversarially robust *input* → Worse calibrated and less stable *prediction*

Google Research

Correlation

- **Adversarial Robustness** (Larger Adv. perturbation ⇨ More Adv. robust input x)
- **Calibration** (Larger ECE ⇨ Worse calibrated prediction $f(x)$)
- **Stability** (Larger variance ⇨ Less stable prediction $f(x)$)



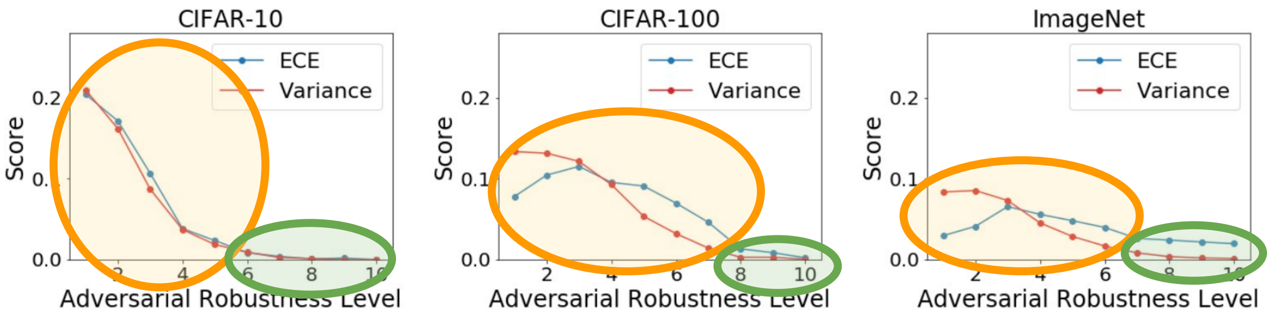
Larger adversarial robustness level → More adv. robust input

Higher adversarially robust input → Better calibrated and more stable prediction

Google Research

Correlation

- **Adversarial Robustness** (Larger Adv. perturbation ⇨ More Adv. robust input x)
- **Calibration** (Larger ECE ⇨ Worse calibrated prediction $f(x)$)
- **Stability** (Larger variance ⇨ Less stable prediction $f(x)$)



Correlation: **Adversarially unrobust** input data are more likely to have **miscalibrated** (higher ECE) and **unstable** (higher variance) predictions.

Can we improve calibration and stability through the relationship with adversarial robustness?

Can we improve calibration and stability through the relationship with adversarial robustness?

To soften the labels of training data based on their adversarial robustness!

Google Research

Can we improve calibration and stability through the relationship with adversarial robustness?

To soften the labels of training data based on their adversarial robustness!

Identify adversarially unrobust training input

Google Research

Can we improve calibration and stability through the relationship with adversarial robustness?

To soften the labels of training data based on their adversarial robustness!

Identify adversarially unrobust training input

Close to decision boundary

Can we improve calibration and stability through the relationship with adversarial robustness?

To soften the labels of training data based on their adversarial robustness!

Identify adversarially unrobust training input

Close to decision boundary

Reduce the confidence of their labels

Google Research

Can we improve calibration and stability through the relationship with adversarial robustness?

To soften the labels of training data based on their adversarial robustness!

Identify adversarially unrobust training input	Close to decision boundary	Reduce the confidence of their labels	Teach the model to be uncertain on Adv. unrobust input

Algorithm

Adversarial **R**obustness based **Ada**ptive **L**abel **S**moothing (**AR-AdaLS**)

- Step 1: Sort and divide the training data into *R=10* small subsets with equal size based on their adversarial robustness

Algorithm

Adversarial **R**obustness based **Ada**ptive **L**abel **S**moothing (**AR-AdaLS**)

- Step 1: Sort and divide the training data into *R=10* small subsets with equal size based on their adversarial robustness
- Step 2: Automatically learn the soft labels in each **training subset** based on calibration performance on the corresponding **validation subset**.

Soft labels for training inputs	**based on**	**Calibration performance** on validation data

Algorithm

Adversarial **R**obustness based **Ada**ptive **L**abel **S**moothing (**AR-AdaLS**)

- Step 1: Sort and divide the training data into *R=10* small subsets with equal size based on their adversarial robustness
- Step 2: Automatically learn the soft labels in each **training subset** based on calibration performance on the corresponding **validation subset**.

$$\text{Update } \widetilde{p}_{r,t}^{z=y} \leftarrow \widetilde{p}_{r,t}^{z=y} - \alpha \cdot \left(\text{conf}(S_r^{val})_t - \text{acc}(S_r^{val})_t\right)$$

Soft label for the correct class in the training subset

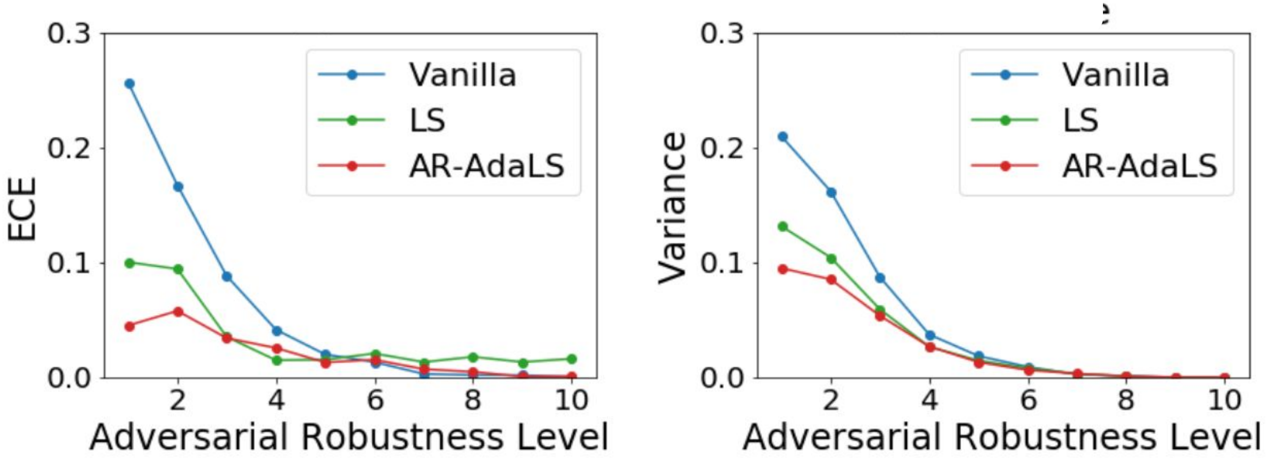
Confidence of the predicted class in validation subset

Accuracy in the validation subset

Google Research

Improvement over Label Smoothing (LS)

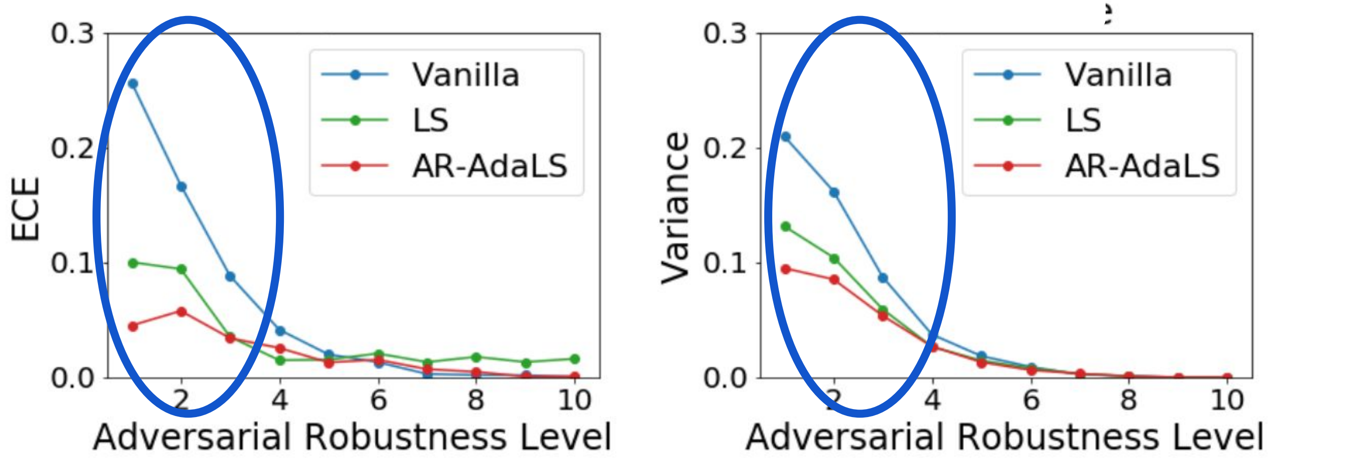
- AR-AdaLS is especially better at improving calibration and stability in **adversarially unrobust regions**, not just on average.



Google Research

Improvement over Label Smoothing (LS)

- AR-AdaLS is especially better at improving calibration and stability in **adversarially unrobust regions**, not just on average.



Compared to existing methods

- AR-AdaLS effectively improves calibration and is only rivaled by domain-knowledge based data augmentation or ensemble models.

Method	CIFAR-10	CIFAR-100	Method	CIFAR-10	CIFAR-100
Single-model based			Data-augmentation based		
Vanilla	2.5	6.1	mixup	0.8	**1.8**
Temperature Scaling	0.8	4.3	CCAT	2.4	4.2
Label Smoothing	1.1	2.8	Ensemble based		
AdaLS	1.3	2.9	Mix-n-Match	1.0	2.8
AR-AdaLS	**0.6**	**2.3**	Ensemble of Vanilla	0.9	**2.2**

Table 1: Expected calibration error (ECE) on CIFAR-10 and CIFAR-100. (Lower ECE is better.)

Improve calibration on shifted dataset

- **Corruptions:** CIFAR-10-C and ImageNet-C include different types of corruptions, e.g., noise, blur, weather and digital categories that frequently encountered in natural images.

Improve calibration on shifted dataset

- **Corruptions:** CIFAR-10-C and ImageNet-C include different types of corruptions, e.g., noise, blur, weather and digital categories that frequently encountered in natural images.
- **Single & Ensemble:**
  - Single AR-AdaLS can effectively improve calibration on shifted data.

Single-model based			Ensemble-based		
Methods	CIFAR-10-C	ImageNet-C	Methods	CIFAR-10-C	ImageNet-C
Vanilla	16.7	12.8	Ensemble of Vanilla	6.5	4.2
LS	10.1	8.2	Ensemble of LS	4.6	4.7
AdaLS	9.6	8.0	Ensemble of AdaLS	5.2	4.8
AR-AdaLS	**6.4**	**6.8**	Ensemble of AR-AdaLS	5.5	5.1
			AR-AdaLS of Ensemble	**4.4**	**4.0**

Table 1: Expected calibration error (ECE) on CIFAR-10-C and ImageNet-C. (Lower ECE is better.)

Improve calibration on shifted dataset

- **Corruptions:** CIFAR-10-C and ImageNet-C include different types of corruptions, e.g., noise, blur, weather and digital categories that frequently encountered in natural images.
- **Single & Ensemble:**
  - Single AR-AdaLS can effectively improve calibration on shifted data.
  - AR-AdaLS can be applied to ensemble models and further improve calibration.

Single-model based			Ensemble-based		
Methods	CIFAR-10-C	ImageNet-C	Methods	CIFAR-10-C	ImageNet-C
Vanilla	16.7	12.8	Ensemble of Vanilla	6.5	4.2
LS	10.1	8.2	Ensemble of LS	4.6	4.7
AdaLS	9.6	8.0	Ensemble of AdaLS	5.2	4.8
AR-AdaLS	**6.4**	**6.8**	Ensemble of AR-AdaLS	5.5	5.1
			AR-AdaLS of Ensemble	**4.4**	**4.0**

Table 1: Expected calibration error (ECE) on CIFAR-10-C and ImageNet-C. (Lower ECE is better.)

Improve stability on shifted dataset

- **Corruptions:** CIFAR-10-C and ImageNet-C include different types of corruptions, e.g., noise, blur, weather and digital categories that frequently encountered in natural images.

Dataset	CIFAR10-C						ImageNet-C					
Shift Intensity	1	2	3	4	5	Mean	1	2	3	4	5	Mean
Vanilla	7.85	9.69	11.2	13.1	16.0	11.6	5.28	6.39	7.37	8.23	8.29	7.11
LS	5.54	6.95	8.11	9.65	11.8	8.41	4.86	5.84	6.78	7.55	7.41	6.49
AdaLS	5.47	6.87	7.95	9.44	11.5	8.25	4.79	5.77	6.66	7.51	7.56	6.46
AR-AdaLS	**4.21**	**5.06**	**5.73**	**6.66**	**8.24**	**5.98**	**4.53**	**5.49**	**6.12**	**6.76**	**6.66**	**5.91**

Table 1: Variance on CIFAR-10-C and ImageNet-C. (Lower variance means more stable.)

Conclusion

- **Relationship among different aspects of robustness**
  - Inputs that are more *vulnerable to adversarial attacks* are more likely to have *poorly calibrated* and *unstable* predictions.
- **AR-AdaLS**
  - Automatically learn how much to soften the labels of training data based on their adversarial robustness.
  - AR-AdaLS can be applied to both single model and ensembles to improve models' calibration and stability.

Thanks!