NOISY COLLABORATION IN KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation is an effective model compression technique in which a smaller model is trained to mimic a larger pretrained model. However in order to make these compact models suitable for real world deployment, not only do we need to reduce the performance gap but also we need to make them more robust to commonly occurring and adversarial perturbations. Noise permeates every level of the nervous system, from the perception of sensory signals to the generation of motor responses. We therefore believe that noise could be a crucial element in improving neural networks training and addressing the apparently contradictory goals of improving both the generalization and robustness of the model. Inspired by trial-to-trial variability in the brain that can result from multiple noise sources, we introduce variability through noise at either the input level or the supervision signals. Our results show that noise can improve both the generalization and robustness of the model. "Fickle Teacher" which uses dropout in teacher model as a source of response variation leads to significant generalization improvement. "Soft Randomization", which matches the output distribution of the student model on the image with Gaussian noise to the output of the teacher on original image, improves the adversarial robustness manifolds compared to the student model trained with Gaussian noise. We further show the surprising effect of random label corruption on a model's adversarial robustness. The study highlights the benefits of adding constructive noise in the knowledge distillation framework and hopes to inspire further work in the area.

1 INTRODUCTION

The design of Deep Neural Networks (DNNs) for efficient real world deployment involves careful consideration of following key elements: memory and computational requirements, performance, reliability and security. DNNs are often deployed in resource constrained devices or in applications with strict latency requirements such as self driving cars which leads to a necessity for developing compact models that generalizes well. Furthermore, since the environment in which the models are deployed are often constantly changing, it is important to consider their performance on both in-distribution data as well as out-of-distribution data. Thereby ensuring the reliability of the models under distribution shift. Finally, the model needs to be robust to malicious attacks by adversaries (Kurakin et al., 2016).

Many techniques have been proposed for achieving high performance in compressed model such as model quantization, model pruning, and knowledge distillation. In our study, we focus on knowledge distillation as an interactive learning method which is more similar to human learning. Knowledge Distillation involves training a smaller network (student) under the supervision of a larger pre-trained network (teacher). In the original formulation, Hinton et al. (2015) proposed mimicking the softened softmax output of the teacher model which consistently improves the performance of the student model compared to the model trained without teacher assistance. However, despite the promising performance gain, there is still a significant performance gap between the student and the teacher model. Consequently an optimal method of capturing knowledge from the larger network and transferring it to a smaller model remains an open question. While reducing this generalization gap is important, in order to truly make these models suitable for real world deployment, it is also pertinent to incorporate methods into the knowledge distillation framework that improve the robustness of the student model to both commonly occurring and malicious perturbations.

For our proposed methods, we derive inspiration from studies in neuroscience on how humans learn. A human infant is born with billions of neurons and throughout the course of its life, the connections between these neurons are constantly changing. This neuroplasticity is at the very core of learning (Draganski et al., 2004). Much of the learning for a child happens not in isolation but rather through collaboration. A child learns by interacting with the environment and understanding it through their own experience as well as observations of others. Two learning theories are central to our approach: cognitive bias and trial-to-trial response variation. Human decision-making shows systematic simplifications and deviations from the tenets of rationality ('heuristics') that may lead to sub-optimal decisional outcomes ('cognitive biases') (Korteling et al., 2018). These biases are strengthened through repeatedly rewarding a particular response to the same stimuli. Trial-to-trial response variation in the brain, i.e. variation in neural responses to the same stimuli, encodes valuable information about the stimuli (Scaglione et al., 2011). We hypothesize that introducing constructive noise in the student-teacher collaborative learning framework to mimic the trial-to-trial response variation in humans can act as a deterrent to cognitive bias which is manifested in the form of memorization and over-generalization in neural networks. When viewed from this perspective, noise can be a crucial element in improving learning and addressing the apparent contradictory goals of achieving accurate and robust models.

In this work, we present a compelling case for the beneficial effects of introduction of noise in knowledge distillation. We provide a comprehensive study on the effects of noise on model generalization and robustness. Our contributions are as follows:

- A comprehensive analysis on the effects of adding a diverse range of noise types in different aspects of the teacher-student collaborative learning framework. Our study aims to motivate further work in exploring how noise can improve both generalization and robustness of the student model.
- A novel approach for transferring teacher model's uncertainty to a student using Dropout in teacher model as a source of trial-to-trial response variability which leads to significant generalization improvement. We call this method "Fickle Teacher".
- A novel approach for using Gaussian noise in the knowledge distillation which improves the adversarial robustness of the student model by an order of magnitude while significantly limiting the drop in generalization. we refer to this method as "Soft Randomization".
- Random label corruption as a strong deterrent to cognitive bias and demonstrating its surprising ability to significantly improve adversarial robustness with minimal reduction in generalization.

2 Related Work

Many experimental and computational methods have reported the presence of noise in the nervous system and how it affects the the function of system (Faisal et al., 2008). Noise as a common regularization technique has been used for ages to improve generalization performance of overparameterized deep neural networks by adding it to the input data, the weights or the hidden units (An, 1996; Steijvers & Grünwald, 1996; Graves, 2011; Blundell et al., 2015; Wan et al., 2013). Many noise techniques have been shown to improve generalization such as Dropout (Srivastava et al., 2014) and injection of noise to the gradient (Bottou, 1991; Neelakantan et al.). Many works show that noise is crucial for non-convex optimization (Zhou et al., 2017; Li & Yuan, 2017; Kleinberg et al., 2018; Yim et al., 2017). A family of randomization techniques that inject noise in the model both during training and inference time are proven to be effective to the adversarial attacks (**?**Xie et al., 2017; Rakin et al., 2018; Liu et al., 2018). Randomized smoothing transforms any classifier into a new smooth classifier that has certifiable l_2-norm robustness guarantees (Lecuyer et al., 2018; Cohen et al., 2019). Label smoothing improves the performance of deep neural networks across a range of tasks (Szegedy et al., 2016; Pereyra et al., 2017). However, Müller et al. (2019) reports that label smoothing impairs knowledge distillation. We believe the knowledge distillation framework with the addition of constructive noise might offer a promising direction towards the design goal mentioned earlier, i.e. achieving lightweight well generalizing models with improved robustness to both adversarial and naturally occurring perturbations.

3 EXPERIMENTAL SETUP

For our empirical analysis, we adopted CIFAR-10 because of its pervasiveness in both knowledge distillation and robustness literature. Furthermore, the size of the dataset allows for extensive experimentation. To study the effect of noise addition in the knowledge distillation framework, we use Hinton method (Hinton et al., 2015) which trains the student model by minimizing the Kullback–Leibler divergence between the smoother output probabilities of the student and teacher model. In all of our experiments we use $\alpha = 0.9$ and $\tau = 4$. We conducted our experiments on Wide Residual Networks (WRN) (Zagoruyko & Komodakis, 2016b). Unless otherwise stated, we normalize the images between 0 and 1 and use standard training scheme as used in (Zagoruyko & Komodakis, 2016a; Tung & Mori, 2019): SGD with momentum; 200 epochs; batch size 128; and an initial learning rate of 0.1, decayed by a factor of 0.2 at epochs 60, 120 and 150. For the choice of student and teacher model architecture, we used WRN-40-2 with 2.2M parameters and WRN-16-2 with 0.7M parameters as the student model. In all of our experiments, we train each model for 5 different seed values. For the teacher model, we select the model with highest test accuracy and then use it to train the student model again for 5 different seed values and report the mean performance.

To evaluate the out of distribution generalization of our models, we used the ImageNet (Krizhevsky et al., 2012) images from the CINIC dataset (Darlow et al., 2018). For adversarial robustness evaluation, we use the Projected Gradient Descent (PGD) attack from Kurakin et al. (2016) and run for multiple step sizes. We report the worst robustness accuracy for 5 random initialization runs. Finally, we test the robustness of our models to commonly occurring corruptions and perturbations proposed by Hendrycks & Dietterich (2019) in CIFAR-C as a proxy for natural robustness. For details of the methods, please see appendex.

4 EMPIRICAL STUDY OF NOISES

In this section, we propose injecting different types of noise in the student-teacher learning framework of knowledge distillation and analyze their effect on the generalization and robustness of the model.

4.1 SIGNAL-DEPENDENT NOISE

Here, we add a signal-dependent noise to the output logits of the teacher model. For each sample, we add zero-mean Gaussian noise with variance that is proportional to the output logits in the given sample (z_i).

$$\hat{z_i} = (1 + \delta).z_i, \quad \delta \sim \mathcal{N}(\mu = 0, \sigma^2) \tag{1}$$

We study the effect for the noise range $[0 - 0.5]$ at steps of 0.1. Figure 1 shows for noise levels up to 0.1, the random signal-dependent noise improves the generalization to CIFAR-10 test set compared to the Hinton method without noise while marginally reducing the out-of-distribution generalization to CINIC-ImageNet. Figure 1 and Figure 11 show a slight increase in the adversarial robustness and natural robustness of the models.

Müller et al. reported that when the teacher model is trained with label smoothing, the knowledge distillation to the student model is impaired and the student model performs worse. On the contrary, for lower level of noise, our method improves the effectiveness of distillation process. Our method differs from their approach in that we train the teacher model without any noise and only when distilling knowledge to the student, we add noise to its softened logits.

4.2 FICKLE TEACHER

Inspired by trial-to-trial variability in the brain and its constructive role in learning, we propose using dropout in the teacher model as a source of variability in the supervision signal from the teacher. We train the teacher model with dropout and while training the student model, we keep the dropout active in the teacher model. As a result, repeated representation of the same input image leads to different output prediction of teacher. Gal & Ghahramani used dropout to obtain principled uncertainty

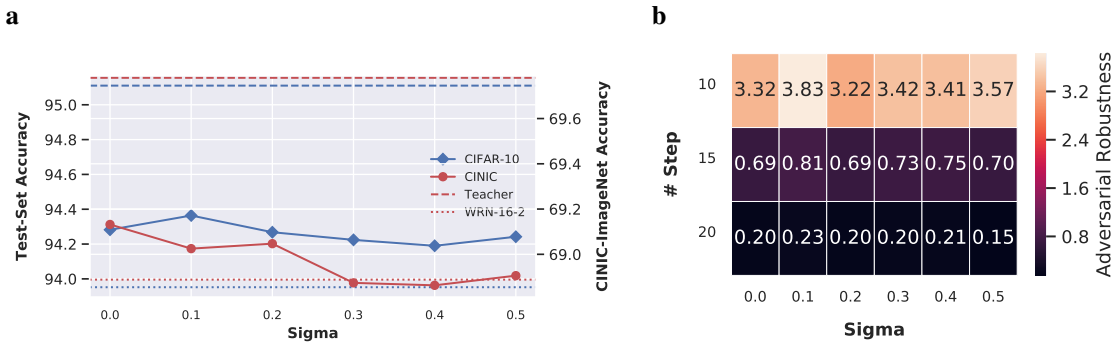

Figure 1: Lower level of signal-dependent Gaussian noise on supervisory signal from teacher improves (**a**) the accuracy of student on the unseen data, but not the generalization to the out-of-distribution data as well as (**b:**) the robustness to PGD attack.

estimates from deep learning networks. Gurau et al. utilize knowledge distillation to better calibrate a student model with the same architecture as the teacher model by using the soft target distribution obtained by averaging the Monte Carlo samples. Our proposed method differs from their method in a number of ways. We use dropout as a source of uncertainty encoding noise for distilling knowledge to a compact student model. Also, instead of averaging Monte Carlo simulations, we used the logits returned by the teacher model with activate dropout and train the student for more epochs so that it can capture the uncertainty of the teacher directly.

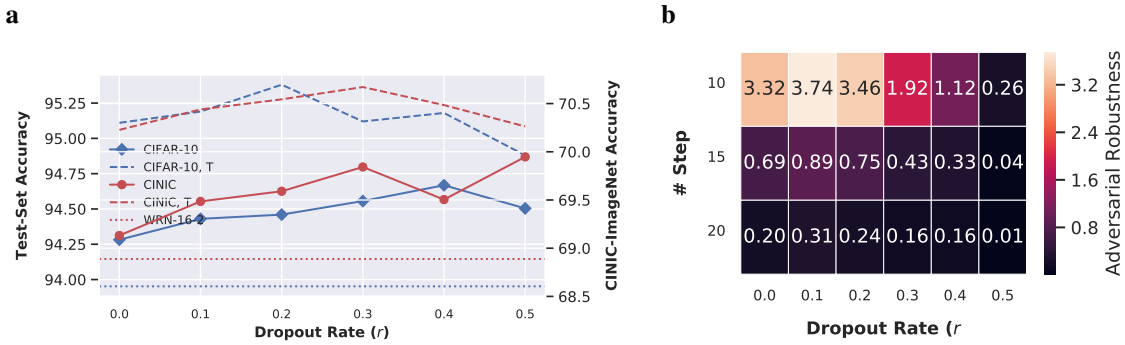

Figure 2: Encoding the uncertainty of teacher helps the student to (**a**)generalize better on both unseen data and out-of-distribution data, and (**b**) to ave higher generalization to PGD attack. Note that for higher dropout rate the performance of teacher drops.

We compare the generalization and robustness of the proposed method for dropout in the range $[0 - 0.5]$ at steps of 0.1. For training parameters, please see the appendex. Figure 12a show that training the student model with dropout using our scheme significantly improves both in-distribution and out-of-distribution generalization over the Hinton method. Interestingly, even when the performance of the teacher model used to train the model is decreasing after drop rate 0.2, the student model performance still improves up to drop rate 0.4. For dropout rate upto 0.2, both PGD Robustness (Figure 12b) and natural robustness increases (Figure 6). This suggest that as per our hypothesis, adding trial-to-trial variability helps in distilling knowledge to the student model.

4.3 Soft Randomization

Pinot et al. provided theoretical evidence for the relation between adversarial robustness and the intensity of random noise injection in the input image. They show that injection of noise drawn from the exponential family such as Gaussian or Laplace noise leads to guaranteed robustness to adversarial attack. However this improved robustness comes at the cost of generalization.

We propose a novel method for adding Gaussian noise in the input image while distilling knowledge to the student model. Since the knowledge distillation framework provides an opportunity to combine multiple sources of information, we hypothesize that using the teacher model trained on clean images, to train the student model with random Gaussian noise can retain the adversarial robustness gain observed with randomized training and mitigate the loss in generalization. Our method involves minimizing the following loss function in the knowledge distillation framework.

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{CE}(S(x+\delta), y) + \alpha\tau^2 D_{KL}(S^\tau(x+\delta)||T^\tau(x)), \quad \delta \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

where $S(.)$ denotes the output of student, $S^\tau(.)$ and $T^\tau(.)$ denote the soften logits of student and teacher models by temperature τ, respectively. α and τ are the balancing factor and temperature parameters from the Hinton method.

We trained the models with six Gaussian noise levels and observe a significant increase in adversarial robustness and a decrease in generalization. However, our proposed method outperforms the compact model trained with Gaussian noise without teacher assistance for both generalization and robustness (Figures: 3 and 4). Our method is able to increase the adversarial robustness even at lower noise intensity For $\sigma = 0.05$, our method achieves 33.85% compared to 3.53% for the student model trained alone. In addition, our method also improves the robustness to common corruptions. Figure 5 shows that the robustness to noise and blurring corruptions improves significantly as the Gaussian noise intensity increases. For weather corruptions, it improves robustness except for fog and frost. Finally for digital corruption except for contrast and saturation, the robustness improves. We also observe changes in the effect at different intensities, for example for frost, the robustness increases at lower noise level and then decreases for higher intensities. Our method allows the use of lower noise intensity for increasing adversarial robustness while keeping the loss in generalization very low compared to other adversarial training methods.

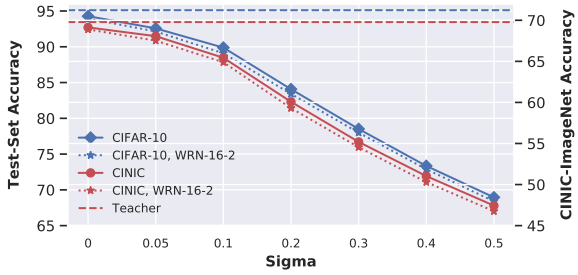

Figure 3: Even adding a small Gaussian noise on input level affects both the accuracy on unseen data and the generalization to out-of-distribution data.

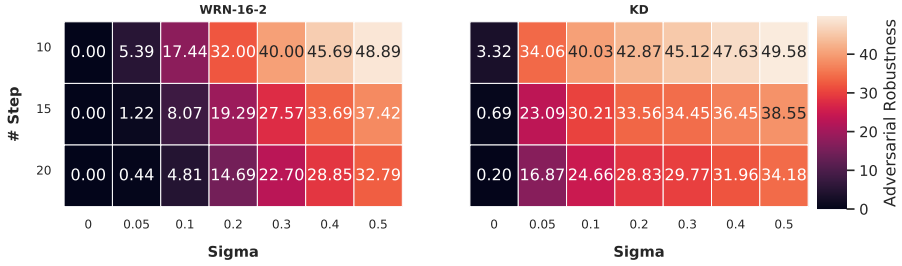

Figure 4: Gaussian noise on input improves the robustness to PGD attack massively.

4.4 Random Label Corruption

Following the analogy with cognitive bias in humans, and relating it to the memorization and over generalization in deep neural networks, we propose a counter intuitive regularization technique based on label noise. For each sample in the training process, with probability p, we randomly change the one hot encoded target labels to an incorrect class. The intuition behind this method is

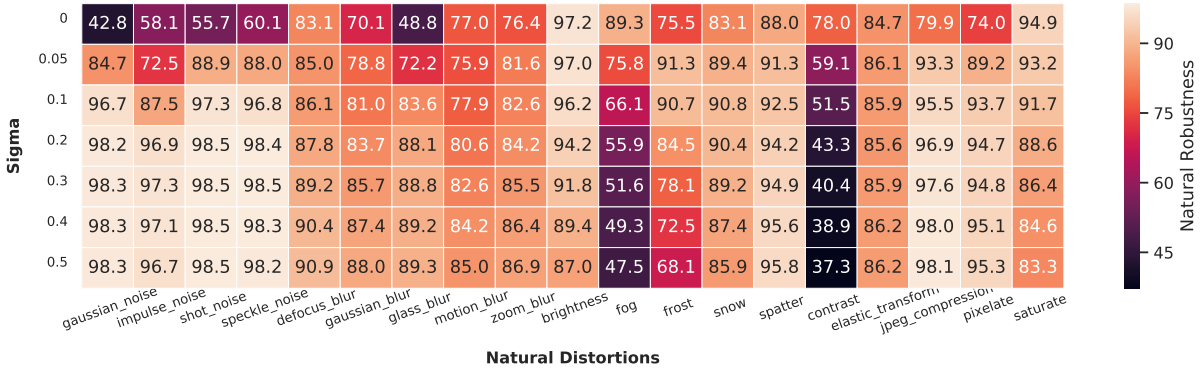

Figure 5: Training student with input corrupted with Gaussian noise improves robustness to most natural distortions.

that by randomly relabeling a fraction of the samples in each epoch, we encourage the model to not be overconfident in its predictions and discourage memorization.

There has been a number of studies on improving the tolerance of the DNNs to noisy labels (Hu et al., 2019; Han et al., 2019; Wang et al., 2019). However, to the best of our knowledge, random label noise has not been explored as a source of constructive noise to improve the generalization of the model.

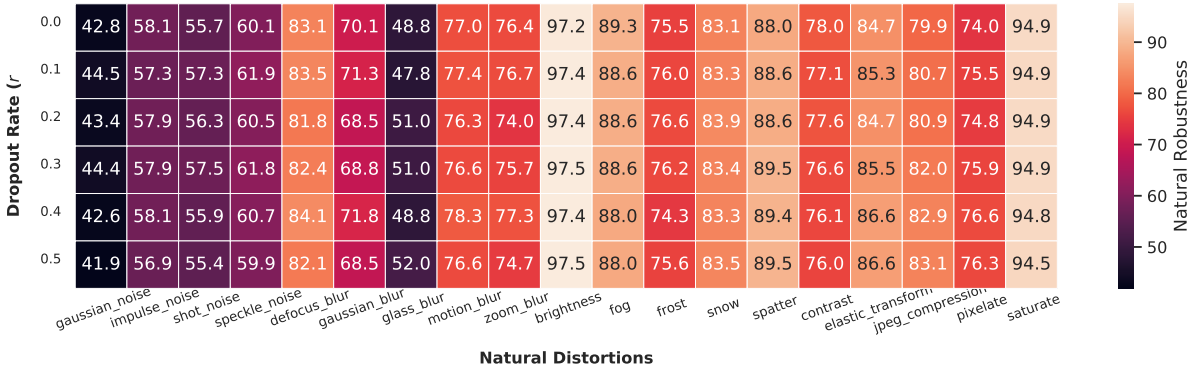

Figure 6: .

We extensively study the effect of random label corruption on a range of p values and at multiple levels: teacher model alone, student model alone, both student and teacher model. When the label corruption is only used during knowledge distillation to student (Corrupted-S), both in-distribution and out-of-distribution generalization increases even for very high corruption levels. When the label corruption is used for training the teacher model and then used to train the student model with (Corrupted-TS) and without (Corrupted-T) label corruption, the generalization drops (Figure 7). In general. knowledge, for high level of label corruption, knowledge distillation outperforms the teacher model. Interestingly, random label corruption leads to a huge increase in adversarial robustness. Just by training with 5% random labels, the PGD-20 robustness of the teacher model increases from 0% to 10.89%. We see this increase in robustness for Corrupted-T and Corrupted-TS. Up to 40% random label corruption, the adversarial robustness increases and slightly decreases for 50%. We believe that this observed phenomenon warrants further study.

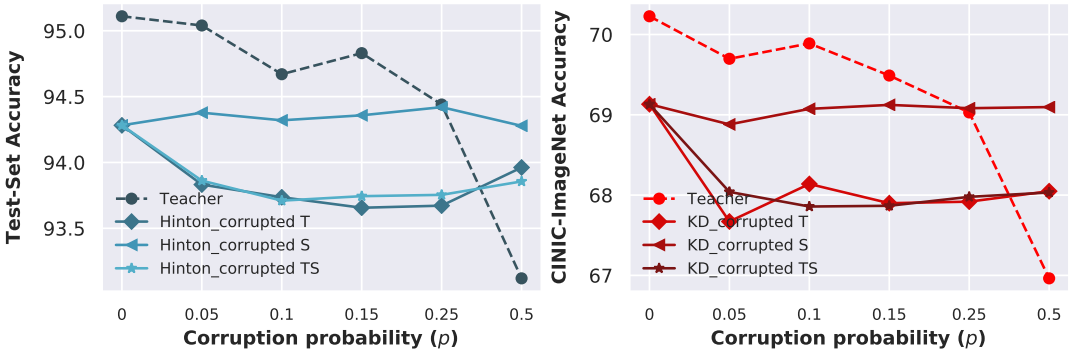

Figure 7: Knowledge distillation of corrupted teacher to both corrupted and clean student decreases the test and generalization accuracy, but from clean teacher to corrupted student the test accuracy improves.

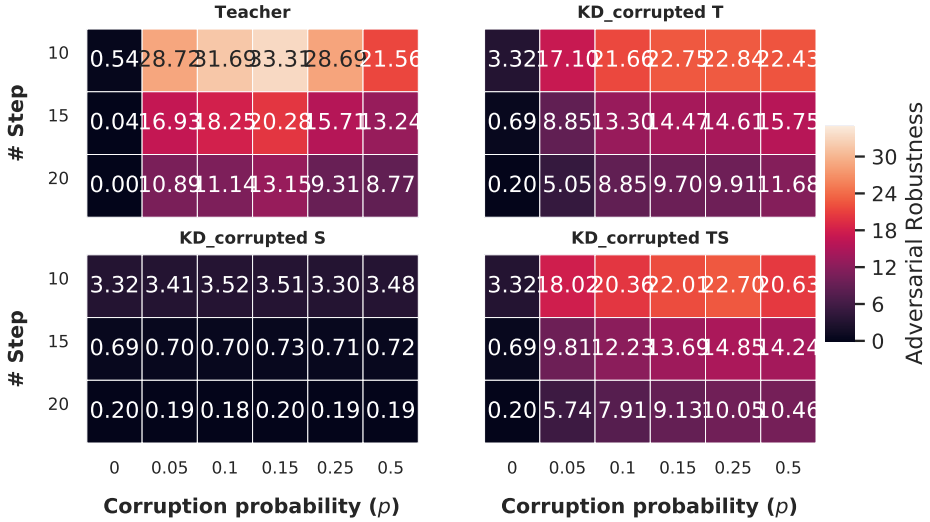

Figure 8: .

5 Conclusion

Inspired by trial-to-trial variability in the brain, we introduce variability in the knowledge distillation framework through noise at either the input level or the supervision signals. For this purpose, we proposed novel ways of introducing noise at multiple levels and studied their effect on both generalization and robustness. Fickle teacher improves the both in-distribution and out of distribution generalization significantly while also slightly improving robustness to common and adversarial perturbations. Soft randomization improves the adversarial robustness of the student model trained alone with Gaussian noise by a huge margin for lower noise intensities while also reducing the drop in generalization. We also showed the surprising effect of random label corruption alone in increasing the adversarial robustness by an order of magnitude in addition to improving the generalization. Our strong empirical results suggest that injecting noises which increase the trial-to-trial variability in the knowledge distillation framework is a promising direction towards training compact models with good generalization and robustness.

References

Guozhong An. The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674, 1996.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Léon Bottou. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8): 12, 1991.

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.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Gavin Weiguang Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, and Ruitong Huang. On the sensitivity of adversarial robustness to input data distributions. *arXiv preprint arXiv:1902.08336*, 2, 2019.

Bogdan Draganski, Christian Gaser, Volker Busch, Gerhard Schuierer, Ulrich Bogdahn, and Arne May. Neuroplasticity: changes in grey matter induced by training. *Nature*, 427(6972):311, 2004.

A Aldo Faisal, Luc PJ Selen, and Daniel M Wolpert. Noise in the nervous system. *Nature reviews neuroscience*, 9(4):292, 2008.

Y Gal and Z Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. arxiv, 2015.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pp. 2348–2356, 2011.

Keren Gu, Brandon Yang, Jiquan Ngiam, Quoc Le, and Jonathan Shlens. Using videos to evaluate image model robustness. *arXiv preprint arXiv:1904.10076*, 2019.

Corina Gurau, Alex Bewley, and Ingmar Posner. Dropout distillation for efficiently estimating model confidence. *arXiv preprint arXiv:1809.10562*, 2018.

Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. *arXiv preprint arXiv:1908.02160*, 2019.

K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. computer vision and pattern recognition (cvpr). In *2016 IEEE Conference on*, volume 5, pp. 6, 2015.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Wei Hu, Zhiyuan Li, and Dingli Yu. Understanding generalization of deep neural networks trained with noisy labels. *arXiv preprint arXiv:1905.11368*, 2019.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.

Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

Johan E Korteling, Anne-Marie Brouwer, and Alexander Toet. A neural network framework for cognitive bias. *Frontiers in psychology*, 9, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. *arXiv preprint arXiv:1802.03471*, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pp. 597–607, 2017.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 369–385, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Jose G Moreno-Torres, Troy Raeder, RocíO Alaiz-RodríGuez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

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

A Neelakantan, L Vilnis, QV Le, I Sutskever, L Kaiser, K Kurach, and J Martens. Adding gradient noise improves learning for very deep networks (2015). *arXiv preprint arXiv:1511.06807*.

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.

Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization: the case of the exponential family. *arXiv preprint arXiv:1902.01148*, 2019.

Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. *arXiv preprint arXiv:1811.09310*, 2018.

Alessandro Scaglione, Karen A Moxon, Juan Aguilar, and Guglielmo Foffani. Trial-to-trial variability in the responses of neurons carries information about stimulus location in the rat whisker thalamus. *Proceedings of the National Academy of Sciences*, 108(36):14956–14961, 2011.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Mark Steijvers and Peter Grünwald. A recurrent network that performs a context-sensitive prediction task. In *Proceedings of the 18th annual conference of the cognitive science society*, pp. 335–339, 1996.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. *arXiv preprint arXiv:1907.09682*, 2019.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pp. 1058–1066, 2013.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. *arXiv preprint arXiv:1908.06112*, 2019.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4133–4141, 2017.

Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *arXiv preprint arXiv:1906.08988*, 2019.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016a.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016b.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn. Stochastic mirror descent in variationally coherent optimization problems. In *Advances in Neural Information Processing Systems*, pp. 7040–7049, 2017.

A APPENDIX

A.1 PRELIMINARIES

In this section we provide details for the methods relevant our study.

A.2 KNOWLEDGE DISTILLATION

Hinton et al. proposed to use the final softmax function with a raised temperature and use the smooth logits of the teacher model as soft targets for the student model. The method involves minimizing the Kullback–Leibler divergence between the smoother output probabilities:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{CE}(\sigma(z_S), y) + \alpha\tau^2 D_{KL}(\sigma(\frac{z_S}{\tau})||\sigma(\frac{z_T}{\tau})) \tag{3}$$

where $\mathcal{L}_{CE}$ denotes cross-entropy loss, $\sigma(.)$ denotes softmax function, z_S student output logit, z_T teacher output logit, τ and α are the hyperparameters which denote temperature and balancing ratio, respectively.

A.3 OUT-OF-DISTRIBUTION GENERALIZATION

Neural networks tend to generalize well when the test data comes from the same distribution as the training data (Deng et al., 2009; He et al., 2015). However, models in the real world often have to deal with some form of domain shift which adversely affects the generalization performance of the models ((Shimodaira, 2000; Moreno-Torres et al., 2012; Kawaguchi et al., 2017; Liang et al., 2017). Therefore, test set performance alone is not the optimal metric for evaluation the generalization of the models in test environment. To measure the out-of-distribution performance, we used the ImageNet (Krizhevsky et al., 2012) images from the CINIC dataset (Darlow et al., 2018). CINIC contains 2100 images randomly selected for each of the CIFAR-10 categories from the ImageNet dataset. Hence the performance of models trained on CIFAR-10 on these 21000 images can be considered as a approximation for a model's out-of-distribution performance.

A.3.1 ADVERSARIAL ROBUSTNESS

Deep Neural Networks have been shown to be highly vulnerable to carefully crafted imperceptible perturbations designed to fool a neural networks by an adversary (Szegedy et al., 2013; Biggio et al., 2013). This vulnerability poses a real threat to deep learning model's deployment in the real world (Kurakin et al., 2016). Robustness to these adversarial attacks has therefore gained a lot of traction in the research community and progress has been to better evaluate robustness to adversarial attacks (Goodfellow et al., 2014; Moosavi-Dezfooli et al., 2016; Carlini & Wagner, 2017) and defend our models against these attacks (Madry et al., 2017; Zhang et al., 2019).

To evaluate the adversarial robustness of models in this study, we use the Projected Gradient Descent (PGD) attack from Kurakin et al. (2016). The PGD-N attack initializes the adversarial image with the original image with the addition of a random noise within some epsilon bound, ϵ. For each step it takes the loss with respect to the input image and moves in the direction of loss with the step size and then clips it within the epsilon bound and the range of valid image.

$$X_0^{adv} = X + U(-\epsilon, +\epsilon) \tag{4}$$

$$X_{N+1}^{adv} = \prod_{\epsilon,d}\{X_N^{adv} + \alpha.sgn(\nabla_X J(X_N^{adv}, y_{true}))\} \tag{5}$$

where ϵ denote epsilon-bound, α step size and X original image. The projection operator $\prod_{\epsilon,d}(A)$ denotes element-wise clipping, with $A_{i,j}$ clipped to the range $[X_{i,j} - \epsilon, X_{i,j} + \epsilon]$ and within valid data range. In all of our experiments, we use 5 random initializations and report the worst adversarial robustness.

A.3.2 NATURAL ROBUSTNESS

While robustness to adversarial attack is important from security perspective, it is an instance of worst case distribution shift. The model also needs to be robust to naturally occurring perturbations which it will encounter frequently in the test environment. Recent works have shown that Deep Neural Networks are also vulnerable to commonly occurring perturbations in the real world which are far from the adversarial examples manifold. Hendrycks et al. (2019) curated a set of real-world, unmodified and naturally occurring examples that causes classifier accuracy to significantly degrade. Gu et al. (2019) measured model's robustness to the minute transformations found across video frames which they refer to as natural robustness and found state-of-the-art classifier to be brittle to these transformations. In their study, they found robustness to synthetic color distortions as a good proxy for natural robustness. In our study we use robustness to the common corruptions and perturbations proposed by Hendrycks & Dietterich (2019) in CIFAR-C as a proxy for natural robustness.

A.3.3 TRADE OFF BETWEEN GENERALIZATION AND ADVERSARIAL ROBUSTNESS

While making our model's robust to adversarial attacks, we need to be careful not to overemphasize robustness to norm bounded perturbation and rigorously test their effect on model's in-distribution and out-of-distribution generalization as well as robustness to naturally occurring perturbation and distribution shift. Recent study have highlighted the adverse affect of adversarially trained model on natural robustness. Ding et al. (2019) showed that even a semantics-preserving transformations on the input data distribution significantly degrades the performance of adversarial trained models but only slightly affects the performance of standard trained model. Yin et al. (2019) showed that adversarially trained models improve robustness to mid and high frequency perturbations but at the expense of low frequency perturbations which are more common in the real world. Furthermore, in the adversarial literature, a number of studies has shown an inherent trade-off between adversarial robustness and generalization Tsipras et al. (2018); Ilyas et al. (2019); Zhang et al. (2019). We would like to point out that these studies were conducted under adversarial setting and do not necessarily hold true for general robustness of the model.

A.4 RANDOM SWAPPING

To exploit the uncertainty of the teacher model for a sample, we propose random swapping noise methods that select a sample with some probability p and then swap the softened softmax logits if the difference is below a threshold.

We propose two variants of random swapping:

1. **Swap Top 2**: Swap the top two logits if the difference between them is below the threshold.
2. **Swap All**: Consider all consecutive pairs iteratively and swap them if the difference is below the threshold value.

These methods improve the in-distribution generalization but adversely affects the out-of-distribution generalization (Figure 9. It does not have a pronounced affect on the robustness (Figures: 9b, 10).

A.5 TRAINING SCHEME FOR DISTILLATION WITH DROPOUT

Because of the variability in the teacher model, the student model needs to be trained to more epochs in order for it to converge and effectively capture the uncertainty of the teacher model.

We used the same initial learning rate of 0.1 and decay factor of 0.2 as per the standard training scheme. For dropout rate of 0.1 and 0.2, we train for 250 epochs and reduce learning rate at 75, 150 and 200 epochs. For dropout rate 0.3, we train for 300 epochs and reduce learning rate at 90, 180 and 240 epochs. Finally for drop rate of 0.4 and 0.5, due to the increased variability, we train for 350 epochs and reduce learning rate at 105, 210 and 280 epochs.

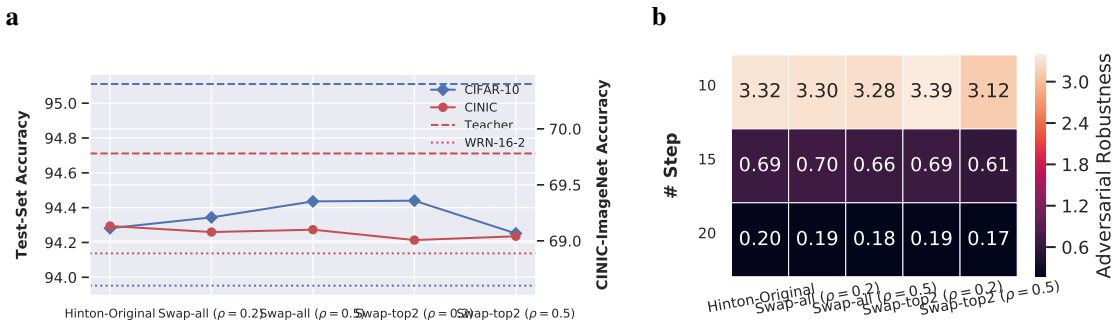

Figure 9: Noise on the supervision from teacher by swapping all logits or the top 2 (**a**) improves the accuracy of student on unseen data, but not the generalization to out-of-distribution data.

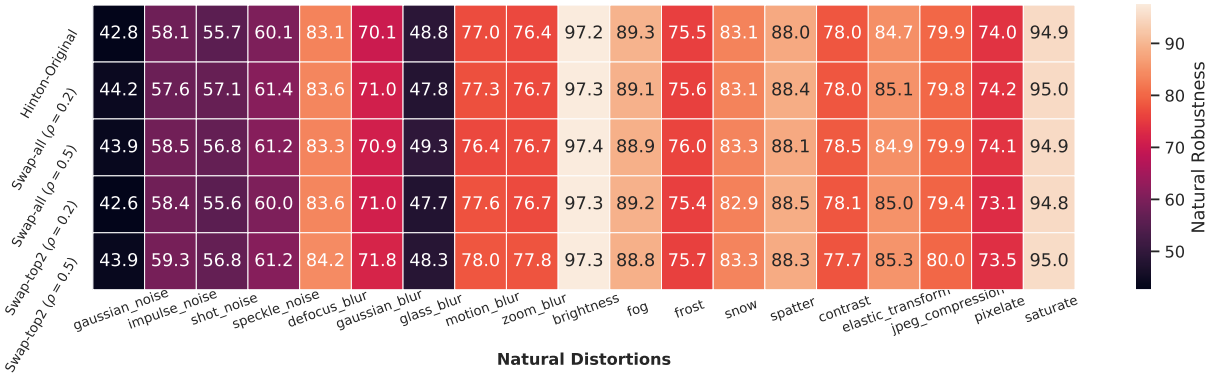

Figure 10: Swapping all logits or the top two if does not improve the robustness to natural distortions, preserves it.

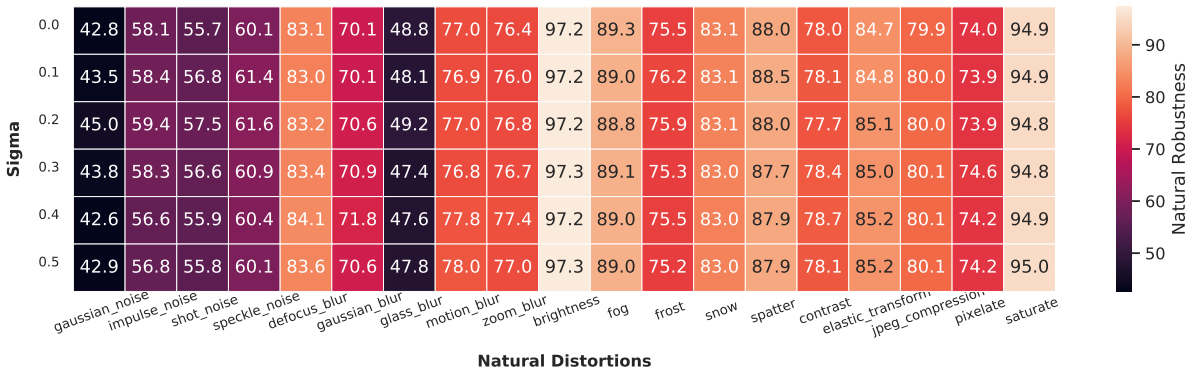

Figure 11: Additive signal-dependent noise maintains the natural robustness to the same level as no noise.

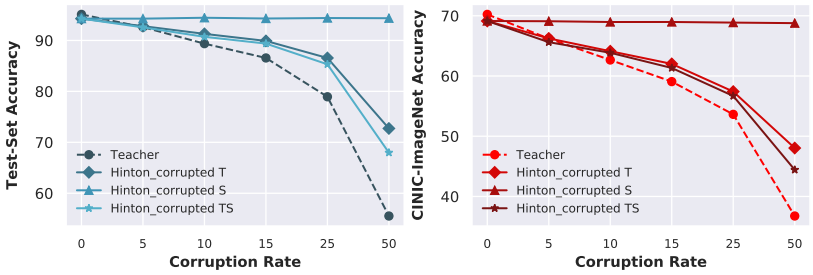

Figure 12: Effect of fix label corruption.

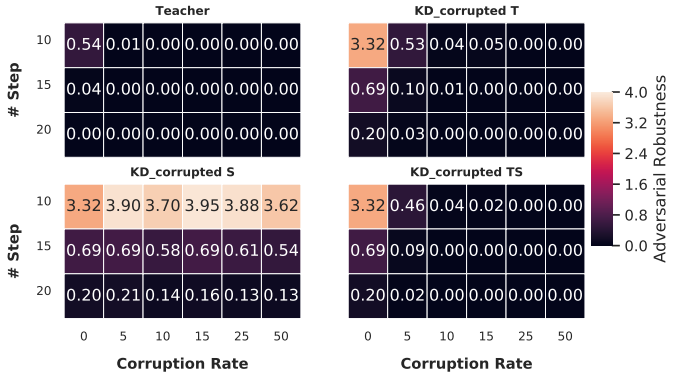

Figure 13: .