IMPROVING GENERALIZATION AND ROBUSTNESS WITH NOISY COLLABORATION IN KNOWLEDGE DISTILLATION

A PREPRINT

Elahe Arani*, Fahad Sarfraz*, and Bahram Zonooz

Advanced Research Lab, NavInfo Europe, Eindhoven, The Netherlands
{elahe.arani, fahad.sarfraz, b.yoosefizonooz}@navinfo.eu

October 14, 2019

ABSTRACT

Inspired by trial-to-trial variability in the brain that can result from multiple noise sources, we introduce variability through noise at different levels in a knowledge distillation framework. We introduce "Fickle Teacher" which provides variable supervision signals to the student for the same input. We observe that the response variability from the teacher results in a significant generalization improvement in the student. We further propose "Soft-Randomization" as a novel technique for improving robustness to input variability in the student. This minimizes the dissimilarity between the student's distribution on noisy data with teacher's distribution on clean data. We show that soft-randomization, even with low noise intensity, improves the robustness significantly with minimal drop in generalization. Lastly, we propose a new technique, "Messy-collaboration", which introduces target variability, whereby student and/or teacher are trained with randomly corrupted labels. We find that supervision from a corrupted teacher improves the adversarial robustness of student significantly while preserving its generalization and natural robustness. Our extensive empirical results verify the effectiveness of adding constructive noise in the knowledge distillation framework for improving the generalization and robustness of the model.

***K*eywords** Knowledge distillation · Generalization · Robustness · Noise

1 Introduction

The design of Deep Neural Networks (DNNs) for efficient real-world deployment involves careful consideration of the 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 autonomous cars which leads to the necessity for developing compact models that generalize well. Furthermore, since the models are often deployed in dynamic environments, it is important to consider their performance on both in-distribution data as well as out-of-distribution data. This ensures the reliability of the models under distribution shift. Finally, the model needs to be robust to malicious attacks by adversaries [32].

A number of techniques have been proposed for achieving high performance in compressed models such as model quantization [61], model pruning [20], and knowledge distillation [24]. Here, we focus on knowledge distillation (KD) as an interactive learning method which is more similar to human learning. Knowledge distillation involves training a smaller model (student) under the supervision of a larger pre-trained model (teacher). In the original formulation, Hinton et al. [24] proposed mimicking the softened softmax output of the teacher which exhibits consistent performance improvement of the student compared to the model trained without teacher assistance. Despite the promising performance gain, there is still a significant generalization gap between student and teacher. Consequently, an optimal method of capturing knowledge from the larger model and transferring it to a smaller model remains an

*Equal contribution

open question. While it is important to reduce the generalization gap, it is also pertinent to incorporate methods into the knowledge distillation framework to make the training and inference of the model more robust.

Noise permeates every level of the nervous system, from the perception of sensory signals to the generation of motor responses [12] and also plays an important role in neural networks training [4, 41]. We therefore hypothesize that noise could help in improving the robustness and generalization of the neural networks. Inspired by trial-to-trial variability in the brain, variations in neural responses to the same stimuli, which can result from multiple noise sources, we introduce variability through noise at input level, supervision signal from teacher or target level in the knowledge distillation framework.

To test our hypothesis, we propose novel ways of injecting noise into the knowledge distillation framework as general and scalable techniques and exhaustively evaluate their effect on generalization and robustness. Our contributions are as follows:

- "Fickel Teacher", a novel approach to simulate trial-to-trial response variability of biological neural networks. The method exposes student to the uncertainty of teacher which results in significant generalization improvement (from 94.28% to 94.67%).
- 'Soft-Randomization", a novel approach for increasing robustness to input variability. The method significantly increases the capacity of the model to learn robust features with small additive noise in the input. For ($\sigma = 0.05$), soft-randomization achieves 16.75% PGD-20 robustness and 92.57% test-set accuracy whereas the model trained with Gaussian data augmentation achieves only 0.41% adversarial robustness and lower generalization (92.14%).
- "Messy-Collaboration", an approach for using target variability as a strong deterrent to cognitive bias. We observe its surprising ability to significantly improve adversarial robustness (from 0.15% to 11.41%) with minimal reduction in generalization (from 94.28% to 93.96%).
- We show the effectiveness of messy-collaboration in learning with noisy labels.
- A comprehensive analysis of the effects of injecting noise in the knowledge distillation framework.

2 Related Work

A number of experimental and computational methods have reported the presence of noise in the nervous system and how it affects the function of the system [12]. Analogously, noise has been used as a common regularization technique to improve the generalization performance of overparameterized deep neural networks by adding it to the input data, the weights or the hidden units [1, 3, 16, 47, 48, 53]. Previous studies also showed that noise is crucial for non-convex optimization [28, 34, 56, 61]. Furthermore, 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 [10, 36, 44, 55]. Randomized smoothing transforms any classifier into a new smooth classifier that has certifiable l_2-norm robustness guarantees [7, 33]. Label smoothing improves the performance of deep neural networks across a range of tasks [42, 49]. However, Müller et al. [40] reports that label smoothing impairs knowledge distillation. On the contrary, we show that our biologically inspired technique of injecting noise into knowledge distillation, *fickle teacher*, significantly improves the generalization of student. Furthermore, *fickle teacher* differs from the works of Bulò et al. [5] and Gurau et al. [18] in that instead of using the soft target distribution obtained by averaging Monte Carlo samples, we use the logits of the teacher model with dropout active directly as a source of uncertainty encoding noise for distilling knowledge to a compact student.

3 Experimental Setup

To study the effect of injecting noise in the knowledge distillation framework, we use Hinton method [24] which trains the student by minimizing the Kullback–Leibler divergence between the smoother output probabilities of the student and teacher. In all of our experiments we use $\alpha = 0.9$ and $\tau = 4$. We conducted our experiments on Wide Residual Networks (WRN) [59]. Unless otherwise stated, we normalize the images between 0 and 1 and use standard training scheme as used in [52, 58]: SGD with 0.99 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. We conducted our experiments on CIFAR-10 [30], with WRN-40-2 with 2.2M parameters as teacher and WRN-16-2 with 0.7M parameters as student. In all of our experiments, we train each model for five different seed values. For the teacher, we select the model with the highest test accuracy and then use it to train the student again for five different seed values and report the mean values for our evaluation metrics.

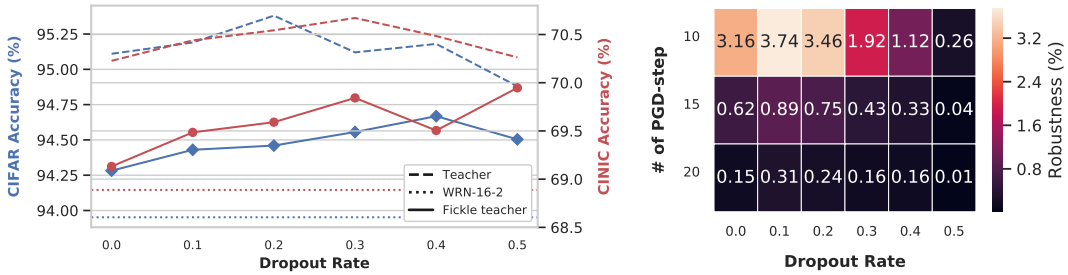

Figure 1: Exposing the student to the uncertainty of the teacher significantly boosts in-distribution (CIFAR-10) and out-of-distribution (CINIC) generalization over the Hinton method (dropout rate = 0; left). Fickle teacher increases the adversarial robustness to PGD attack only for low dropout rate (right).

To approximate the out-of-distribution generalization of our models, we use the ImageNet [31] images from the CINIC dataset [8]. For the evaluation of adversarial robustness, we use the Projected Gradient Descent (PGD) attack from Kurakin et al. [32] and conduct the attack for multiple step sizes. We report the worst robustness accuracy for five random initialization runs. Finally, we test the robustness of our models to commonly occurring corruptions and perturbations proposed by Hendrycks and Dietterich [22] in CIFAR-C as a proxy for natural robustness. We evaluate average robustness to the 19 distortions across 5 severity levels. Furthermore, we calculate mean Corruption Accuracy (mCA) over the 19 distortions. For both adversarial and natural robustness evaluation, we use a fine-grained measure of robustness, by computing accuracy on the corrupted image conditional on the clean image being classified correctly (for details of the methods, see Appendix).

4 Empirical study of Noises

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

4.1 Fickle Teacher

Trial-to-trial response variability in the brain can be considerably different across stimuli, suggesting that it could also provide an important contribution to the information conveyed by the neural responses about the stimuli [45]. Similarly, in deep neural networks, dropout which randomly switches off a group of selected hidden units results in response variability and can be used to obtain principled uncertainty estimates [14]. We, therefore, propose to use dropout in the teacher model to simulate trial-to-trial variability. We first train the teacher with dropout and keep dropout active in the teacher while distilling knowledge to the student. As a result, the teacher provides variable supervision signals to the student for the same input, thereby exposing the student to its uncertainty. We systematically change the dropout rate $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and show its effect on generalization and robustness (for training details, see the Appendix). The results show that fickle teacher significantly improves both in-distribution and out-of-distribution generalization of the student compared to Hinton method (Figure 1, left). Interestingly, for CIFAR-10, even when the accuracy of the teacher decreases after a dropout rate of 0.2, the student accuracy still improves up to a dropout rate of 0.4. For the same student and teacher network architectures, fickle teacher achieves higher performance on CIFAR-10 than the state-of-the-art knowledge distillation methods [52, 58]. For CINIC, as we increase the dropout rate, the student generalization gets closer to the teacher and is the highest at dropout rate 0.5. The adversarial robustness increases for dropout rate up to 0.2 and then decreases (Figure 1, right) while the mCA for natural robustness is maintained or marginally increases with the dropout rate (Figure 2). The results show the effectiveness of fickle teacher in improving the generalization of student.

4.2 Soft-Randomization

Pinot et al. [43] 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 adversarial robustness comes at the cost of generalization. Therefore, an important consideration for methods proposed to increase the adversarial robustness is to reduce the loss in generalization. Since knowledge distillation provides an opportunity to combine multiple sources of information, we hypothesize that combining information from a teacher with high generalization while training the student to be robust to noisy input can reduce the trade-off. To test the hypothesis, we propose a novel technique for

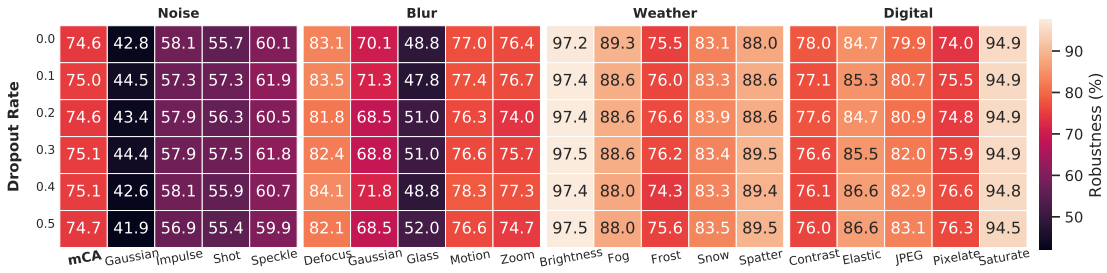

Figure 2: Fickle teacher consistently increases mCA compared to the Hinton method. Lower dropout rate improves robustness to majority of the distortions.

improving robustness to input variability in the student which uses the teacher trained on clean data, to train the student on noisy data. Here, we minimize the dissimilarity between the student's distribution on noisy data with the teacher's distribution on clean data (Figure 3). The loss function is as follows:

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

where $\delta \sim \mathcal{N}(0, \sigma^2)$ is white Gaussian noise. α and τ are the balancing factor and temperature parameter from the Hinton method. $S(.)$ denotes the softmax output of student, $S^\tau(.)$ and $T^\tau(.)$ denote the student's and teacher's softmax output with raised temperature respectively.

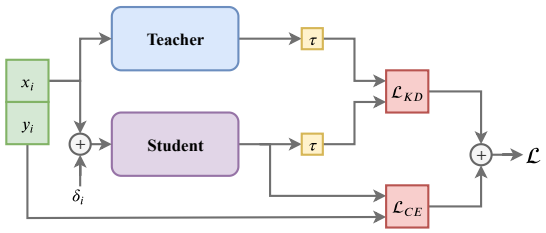

Figure 3: Soft-randomization minimizes the KL divergence between the student response on noisy data and the teacher response on the clean data in addition to minimizing the cross entropy loss on the noisy data.

We train soft-randomization for both low noise intensities $\{0.01, 0.02, 0.03, 0.04, 0.05\}$ and higher noise intensities $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. As a baseline, we compare soft-randomization to the compact model trained alone with Gaussian data augmentation. Figure 4 shows that soft-randomization consistently achieves both higher adversarial robustness and generalization for all σ values compared to Gaussian augmentation. For lower noise intensities, soft-randomization significantly outperforms Gaussian augmentation. For $\sigma = 0.05$, soft-randomization achieves 16.75% robustness to PGD-20 attack and 92.57% CIFAR-10 generalization compared to 0.41% robustness and 92.14% generalization for Gaussian augmentation. Hence, soft-randomization significantly increases the the capacity of the student to learn robust features even with lower noise intensities.

Figure 5 shows the mCA consistently improves over the Hinton method for all Gaussian noise levels. While robustness drops most notably for color distortions (brightness, fog, contrast, and saturation), robustness to noise and blurring corruptions improves significantly as the Gaussian noise intensity increases. 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. Soft-randomization allows the use of lower noise intensity for increasing adversarial robustness while keeping the loss in generalization lower compared to Gaussian augmentation. This provides greater flexibility in finding a suitable trade-off between generalization and robustness based on the application.

4.3 Messy-collaboration

Human decision-making shows systematic simplifications and deviations from the tenets of rationality ('heuristics') which may lead to sub-optimal decisional outcomes ('cognitive biases') [29]. We believe, this cognitive bias is manifested in deep neural networks in the form of memorization and over-generalization, and propose a counter-intuitive regularization technique based on label noise to mitigate cognitive bias. We term this technique as messy-collaboration (MC). For each sample in the training process, with rate r, we randomly change the one-hot encoded target labels to an

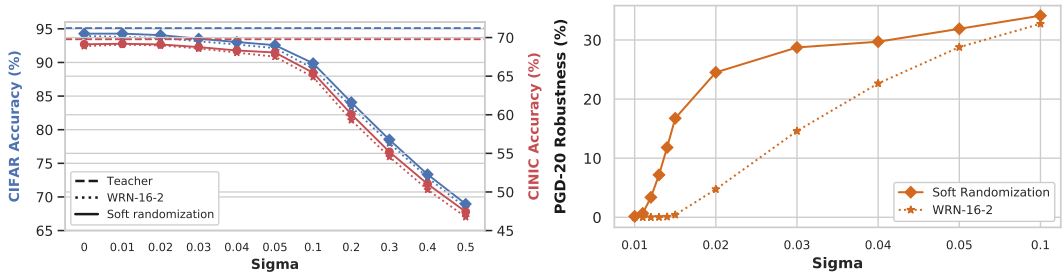

Figure 4: Soft-randomization consistently achieves higher generalization (upper panel) and adversarial robustness (lower panel) compared to Gaussian Augmentation. For lower noise intensities, soft-randomization provides significantly higher robustness to PGD-20 attack.

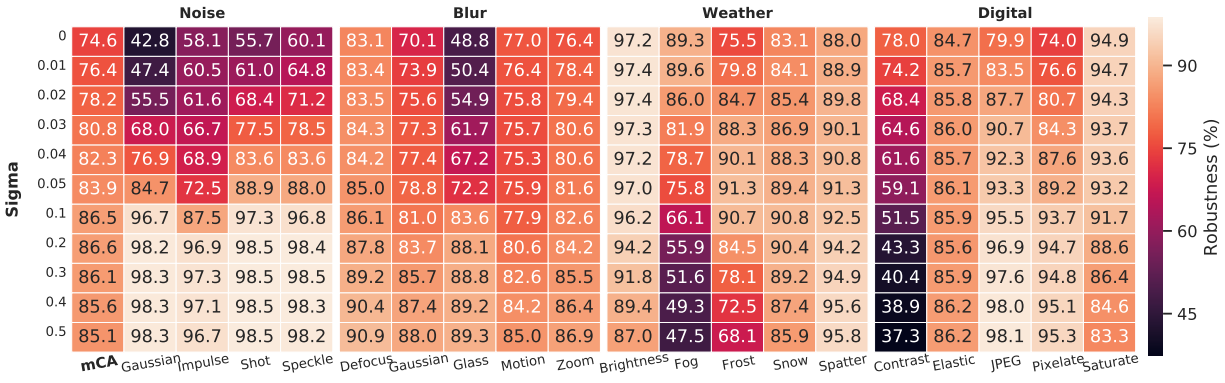

Figure 5: Soft-randomization consistently improves the average robustness to natural distortions (mCA), although the robustness to color distortions is impaired. Most notably, it improves the robustness to noise and blurring distortions.

incorrect class. The intuition behind this method is that by randomly relabeling a fraction of the samples in each batch, we introduce target variability[2] which encourages the model not to be overconfident in its predictions and discourages memorization. There are a number of studies on improving the tolerance of DNNs to noisy labels [19, 25, 54]. However, to the best of our knowledge, random label noise has not been explored as a source of constructive noise for cognitive bias mitigation.

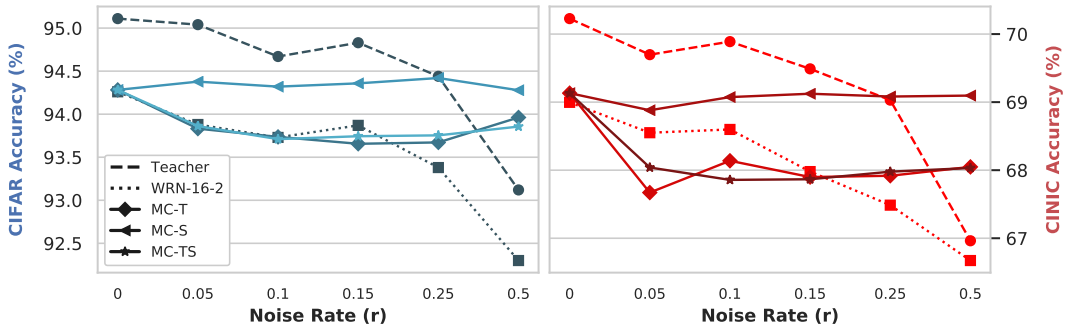

Figure 6: When target variability is introduced only during knowledge distillation to student (MC-S), the generalization increases over Hinton method. However, generalization marginally decreases when the student is trained with a teacher model trained with target variability (MC-T and MC-TS). For higher values of r, the messy-collaboration outperforms teacher trained with target variability.

[2]We use the term *target variability* to refer to the random label corruption which we are introducing intentionally for each batch during training. Whereas, *noisy labels* refers to the inherent corruption in the labels which comes from incorrect annotations.

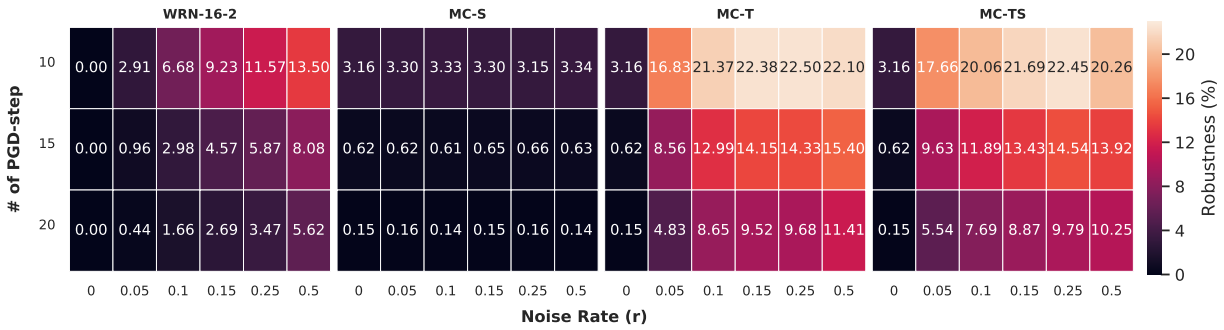

Figure 7: Messy-collaboration increases the robustness to PGD attack (WRN-16-2). The increase in adversarial robustness is more pronounced with messy-collaboration with a teacher trained with target variability (MC-T and MC-TS). However, there is only a marginal effect on adversarial robustness when target variability is only introduced during student training (MC-S).

Our experiments set out to dissociate the effect of injecting target variability in different stages of the knowledge distillation framework. Therefore, we study the effect of systematically increasing the noise rate r for different setups: introducing target variability while training teacher (MC-T), introducing target variability while distilling knowledge to the student (MC-S), and introducing target variability when training the teacher as well as during knowledge distillation to the student (MC-TS).

When target variability is used only during knowledge distillation to the student (MC-S), both in-distribution and out-of-distribution generalization increase over the Hinton method, even for very high noise rates. When target variability is used for training the teacher which is then used to distill knowledge to the student (MC-T and MC-TS), the generalization drops (Figure 6). At higher noise rates ($r = 0.5$), all variants of messy-collaboration significantly outperform the teacher trained with target variability at the same rate. Interestingly, target variability has the remarkable effect on increasing the adversarial robustness. Figure 7 shows that the robustness of the student increases as we increase the noise rate. The increase in adversarial robustness is more pronounced when the teacher model trained with target variability is used to train the student (MC-T and MC-TS). Furthermore, messy-collaboration maintains the natural robustness of Hinton method (Figure 16 in appendix).

To understand the effect of target variability on generalization, we visualize the distribution of the magnitude of softmax probabilities. Figure 8 shows that as we increase the noise rate in messy-collaboration, it not only leads to a smoother output distribution, but also shifts the distribution from overconfident softmax probabilities to lower probabilities. The results show that target variability in messy-collaboration enforces the model not to be overconfident in its predictions and discourages memorization which results in better generalization.

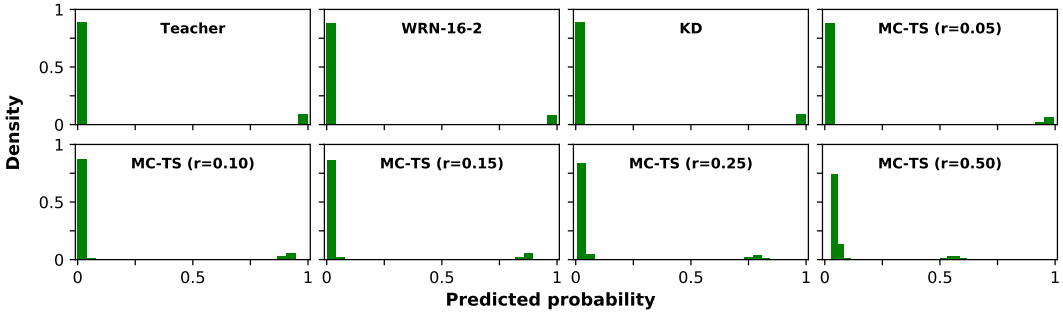

Figure 8: Distribution of the magnitude of softmax probabilities on CIFAR-10 test set. Training the teacher, the compact model alone (WRN-16-2), and using Hinton method to train the student on clean data (KD) lead to a softmax distribution where probabilities are either 0 or 1. By contrast, messy-collaboration (MC-TS) leads to smoother output distributions, which results in better generalization.

4.3.1 Messy-Collaboration for learning with noisy labels

To utilize the vast amount of open-source data available, researchers have proposed methods to generate labels automatically using user tags and keywords. However, these techniques lead to noisy labels which affect the generalization of the model [13]. Considering the abundance of noisy labels, it is important to develop methods that can effectively learn

from noisy labels. Here, we show the effectiveness of knowledge distillation in learning with noisy labels at varying rates of label corruption on CIFAR-10. We further show that messy-collaboration improves the generalization over Hinton method.

Figure 9 shows that the generalization drops with increasing the corruption rate (cf. Teacher and WRN-16-2). For label corruption rate 0.1 and higher, knowledge distillation improves the generalization performance significantly and even outperforms the teacher. The gain in generalization with Hinton method is higher as we increase the label corruption rate. Figure 10 shows the effect of varying the messy-collaboration's noise rate for the different label corruption rates. For label corruption rate over 0.1, messy-collaboration improves the generalization over the student and teacher for all noise rates. Interestingly, at a higher noise rate of 0.5, messy-collaboration improves the generalization over Hinton for all corruption rates. This shows that the target variability in messy-collaboration makes the model more tolerant to label noise which allows efficient learning with noisy labels.

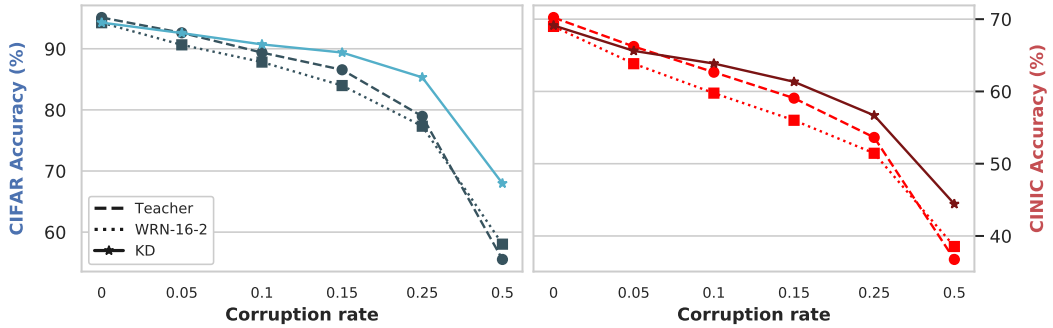

Figure 9: Generalization of WRN-16-2 and teacher drops with increasing corruption rate. Knowledge distillation significantly improves the generalization even over the teacher. The gap between the student and the teacher increases at larger values for corruption rate.

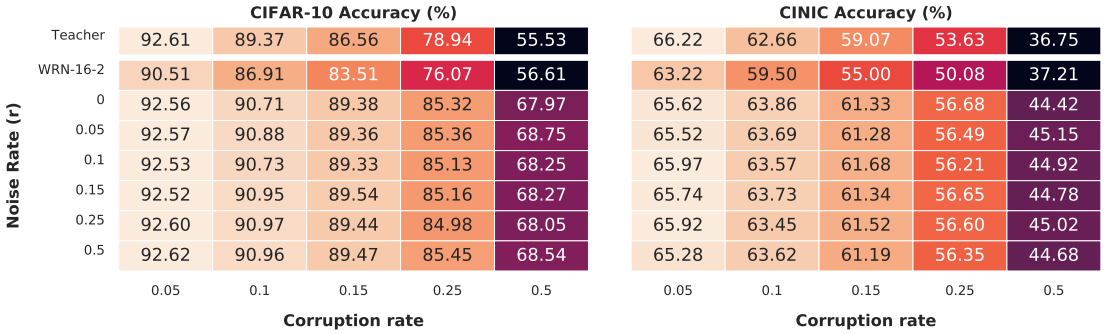

Figure 10: Messy-collaboration improves in-distribution and out-of-distribution generalization over the teacher and the Hinton method in the presence of corrupted labels. At higher corruption rate, the improvement over the Hinton method increases.

5 Conclusion

We proposed novel ways of injecting noise in the knowledge distillation framework and rigorously studied their effect on the generalization and robustness of the model. We introduced *fickle teacher* which exposes the student to its uncertainty using dropout. We show that the variability in the supervision signal improves both in-distribution and out-of-distribution generalization significantly while marginally improving robustness to common and adversarial perturbations. We further proposed *soft-randomization* for increasing the robustness to input variability by matching the output distribution of student on noisy data to the output distribution of teacher on clean data. This significantly increases the capacity of the student to learn robust features. We showed it improves the adversarial robustness of student compared to the model trained alone with Gaussian data augmentation by a significant margin for lower noise intensities, while also reducing the drop in generalization. Finally, we introduced *messy-collaboration* which imposes target variability in different stages of the knowledge distillation framework. We showed the intriguing effect of target variability on increasing the adversarial robustness by an order of magnitude in addition to improving the generalization.

We further showed the effectiveness of messy collaboration in learning with noisy labels. The extensive empirical results demonstrate that injecting noises which increase the trial-to-trial variability in the knowledge distillation framework is a promising direction towards training compact models with improved generalization and robustness.

References

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

[2] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. (2013). Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer.

[3] Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*.

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

[5] Bulò, S. R., Porzi, L., and Kontschieder, P. (2016). Dropout distillation. In *International Conference on Machine Learning*, pages 99–107.

[6] Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE.

[7] Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*.

[8] Darlow, L. N., Crowley, E. J., Antoniou, A., and Storkey, A. J. (2018). Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*.

[9] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[10] Dhillon, G. S., Azizzadenesheli, K., Lipton, Z. C., Bernstein, J., Kossaifi, J., Khanna, A., and Anandkumar, A. (2018). Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*.

[11] Ding, G. W., Lui, K. Y. C., Jin, X., Wang, L., and Huang, R. (2019). On the sensitivity of adversarial robustness to input data distributions. *arXiv preprint arXiv:1902.08336*, 2.

[12] Faisal, A. A., Selen, L. P., and Wolpert, D. M. (2008). Noise in the nervous system. *Nature reviews neuroscience*, 9(4):292.

[13] Frénay, B. and Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

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

[15] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[16] Graves, A. (2011). Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356.

[17] Gu, K., Yang, B., Ngiam, J., Le, Q., and Shlens, J. (2019). Using videos to evaluate image model robustness. *arXiv preprint arXiv:1904.10076*.

[18] Gurau, C., Bewley, A., and Posner, I. (2018). Dropout distillation for efficiently estimating model confidence. *arXiv preprint arXiv:1809.10562*.

[19] Han, J., Luo, P., and Wang, X. (2019). Deep self-learning from noisy labels. *arXiv preprint arXiv:1908.02160*.

[20] Han, S., Mao, H., and Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

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

[22] Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

[23] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. (2019). Natural adversarial examples. *arXiv preprint arXiv:1907.07174*.

[24] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[25] Hu, W., Li, Z., and Yu, D. (2019). Understanding generalization of deep neural networks trained with noisy labels. *arXiv preprint arXiv:1905.11368*.

[26] Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. (2019). Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*.

[27] Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. (2017). Generalization in deep learning. *arXiv preprint arXiv:1710.05468*.

[28] Kleinberg, R., Li, Y., and Yuan, Y. (2018). An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*.

[29] Korteling, J. E., Brouwer, A.-M., and Toet, A. (2018). A neural network framework for cognitive bias. *Frontiers in psychology*, 9.

[30] Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Technical report, Citeseer.

[31] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[32] Kurakin, A., Goodfellow, I., and Bengio, S. (2016). Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.

[33] Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. (2018). Certified robustness to adversarial examples with differential privacy. *arXiv preprint arXiv:1802.03471*.

[34] Li, Y. and Yuan, Y. (2017). Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607.

[35] Liang, S., Li, Y., and Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

[36] Liu, X., Cheng, M., Zhang, H., and Hsieh, C.-J. (2018). Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–385.

[37] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

[38] Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.

[39] Moreno-Torres, J. G., Raeder, T., Alaiz-RodríGuez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

[40] Müller, R., Kornblith, S., and Hinton, G. (2019). When does label smoothing help? *arXiv preprint arXiv:1906.02629*.

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

[42] Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., and Hinton, G. (2017). Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

[43] Pinot, R., Meunier, L., Araujo, A., Kashima, H., Yger, F., Gouy-Pailler, C., and Atif, J. (2019). Theoretical evidence for adversarial robustness through randomization: the case of the exponential family. *arXiv preprint arXiv:1902.01148*.

[44] Rakin, A. S., He, Z., and Fan, D. (2018). Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. *arXiv preprint arXiv:1811.09310*.

[45] Scaglione, A., Moxon, K. A., Aguilar, J., and Foffani, G. (2011). 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.

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

[47] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

[48] Steijvers, M. and Grünwald, P. (1996). A recurrent network that performs a context-sensitive prediction task. In *Proceedings of the 18th annual conference of the cognitive science society*, pages 335–339.

[49] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

[50] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

[51] Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. (2018). Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*.

[52] Tung, F. and Mori, G. (2019). Similarity-preserving knowledge distillation. *arXiv preprint arXiv:1907.09682*.

[53] Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., and Fergus, R. (2013). Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066.

[54] Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., and Bailey, J. (2019). Symmetric cross entropy for robust learning with noisy labels. *arXiv preprint arXiv:1908.06112*.

[55] Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. (2017). Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*.

[56] Yim, J., Joo, D., Bae, J., and Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141.

[57] Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D., and Gilmer, J. (2019). A fourier perspective on model robustness in computer vision. *arXiv preprint arXiv:1906.08988*.

[58] Zagoruyko, S. and Komodakis, N. (2016a). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

[59] Zagoruyko, S. and Komodakis, N. (2016b). Wide residual networks. *arXiv preprint arXiv:1605.07146*.

[60] Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*.

[61] Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S., and Glynn, P. W. (2017). Stochastic mirror descent in variationally coherent optimization problems. In *Advances in Neural Information Processing Systems*, pages 7040–7049.

A Appendix

A.1 Preliminaries

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

A.1.1 Knowledge Distillation

Hinton et al. [24] proposed to use the final softmax function with a raised temperature and use the smooth logits of the teacher as soft targets for the student. The method involves minimizing the Kullback–Leibler divergence between the smoother output probabilities:

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

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.1.2 Out-of-Distribution Generalization

Neural networks tend to generalize well when the test data comes from the same distribution as the training data [9, 21]. 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 [27, 35, 39, 46]. Therefore, test set performance alone is not the optimal metric for evaluating the generalization of the models in test environment. To measure the out-of-distribution performance, we use the ImageNet images from the CINIC dataset [8]. 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 generalization.

A.1.3 Adversarial Robustness

Deep Neural Networks have been shown to be highly vulnerable to carefully crafted imperceptible perturbations designed to fool a neural network by an adversary [2, 50]. This vulnerability poses a real threat to deep learning model's deployment in the real world [32]. Robustness to these adversarial attacks has therefore gained a lot of traction in the research community and progress has been made to better evaluate robustness to adversarial attacks [6, 15, 38] and defend the models against these attacks [37, 60].

To evaluate the adversarial robustness of models in this study, we use the Projected Gradient Descent (PGD) attack from Kurakin et al. [32]. 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.

$$X_0^{adv} = X + U(-\epsilon, +\epsilon) \tag{3}$$

$$X_{N+1}^{adv} = \prod_{\epsilon,d}\{X_N^{adv} + \alpha.sgn(\nabla_X J(X_N^{adv}, y_{true}))\} \tag{4}$$

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 conduct a 10, 15 and 20 step PGD attack with 0.031 epsilon with 0.03 step size and 5 random initializations.

A.1.4 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. [23] curated a set of real-world, unmodified and naturally occurring examples that causes classifier accuracy to significantly degrade. Gu et al. [17] 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 our study we use robustness to the common corruptions and perturbations proposed by Hendrycks and Dietterich [22] in CIFAR-C as a proxy for natural robustness.

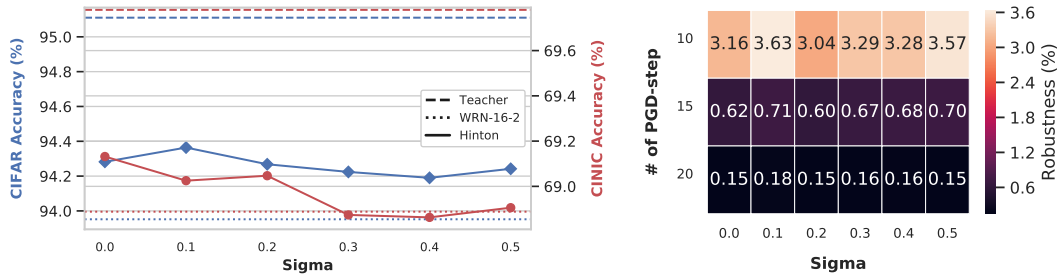

Figure 11: Lower level of signal-dependent noise on supervisory signal from teacher improves the in-distribution generalization and adversarial robustness but not the out-of-distribution generalization.

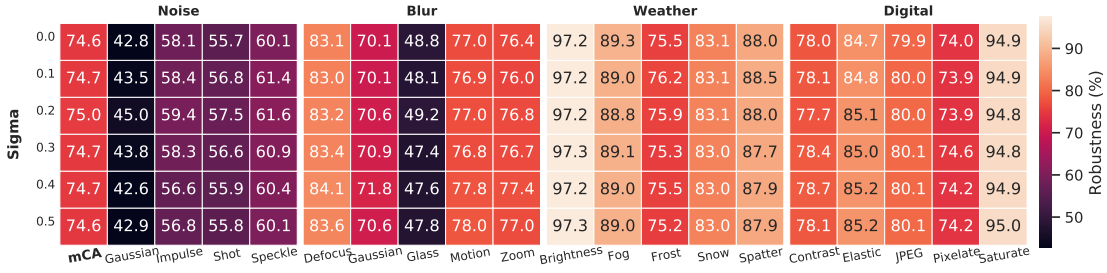

Figure 12: Signal-dependent noise maintains the natural robustness of the Hinton method.

A.1.5 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. We also need to rigorously test the effect of methods designed to increase adversarial robustness 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. [11] 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. [57] 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 [26, 51, 60].

A.2 Additional Experiments

A.2.1 Signal-dependent Noise

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

$$\hat{z}_i = (1 + \delta).z_i, \quad \delta \sim \mathcal{N}(\mu = 0, \sigma^2) \tag{5}$$

We systematically study the effect of increasing the intensity of signal-dependent noise $(0, 0.1, 0.2, 0.3, 0.4, 0.5)$. Figure 11 shows that for noise levels up to 0.1, the random signal-dependent noise improves the in-distribution generalization over the Hinton method. However, it impairs the out-of-distribution generalization to CINIC-ImageNet. Figure 11 and Figure 12 show a slight increase in the adversarial robustness and natural robustness of the models.

Müller et al. [40] reported that when the teacher is trained with label smoothing, the knowledge distillation to the student is impaired and the student performs worse. On the contrary, for lower level of noise, our method improves the effectiveness of knowledge distillation process. Signal-dependent noise differs from their approach in that we train the teacher without any noise and only when distilling knowledge to the student, we add noise to its softened logits.

A.2.2 Random Swapping

To exploit the uncertainty of the teacher for a sample, we inject noise by randomly swapping softened softmax logits. With probability p, we swap the 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 sorted by the logit value and swap the pairs if the difference is below the threshold value.

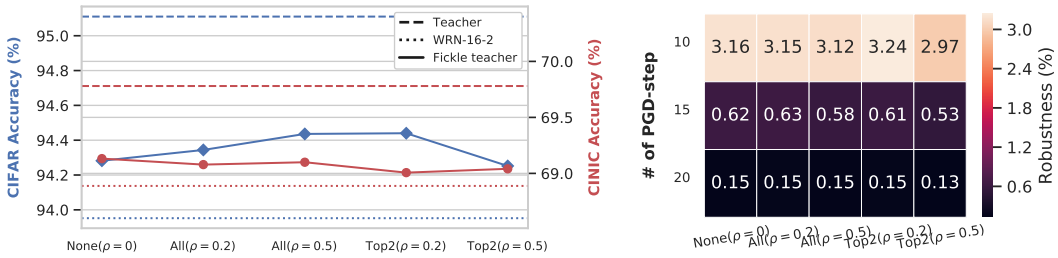

Figure 13: Swapping the logits based on the relative logit difference improves in-distribution generalization, but impairs out-of-distribution generalization (left). Robustness to PGD-20 is maintained except for swap Top2 with higher swapping probability (right).

These technique improves the in-distribution generalization but adversely affects the out-of-distribution generalization (Figure 13. It does not have a significant affect on the adversarial robustness (Figure 13), but it marginally improves mCA (Figure 14).

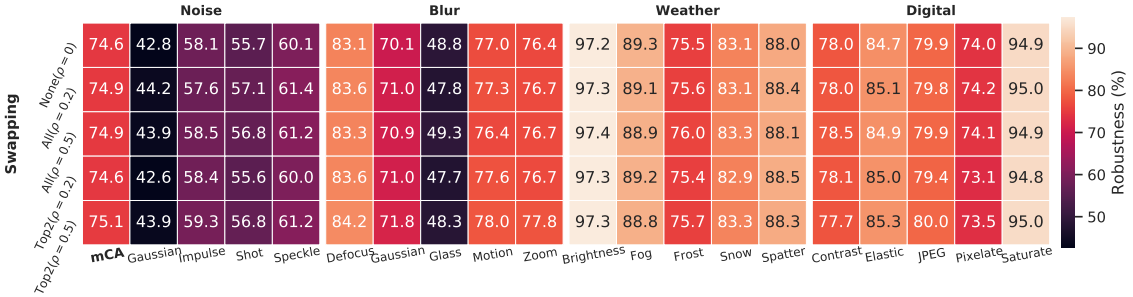

Figure 14: Swapping logits marginally improves the robustness to natural robustness (mCA).

A.3 Training Scheme for Fickle Teacher

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

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.

A.4 Additional Results

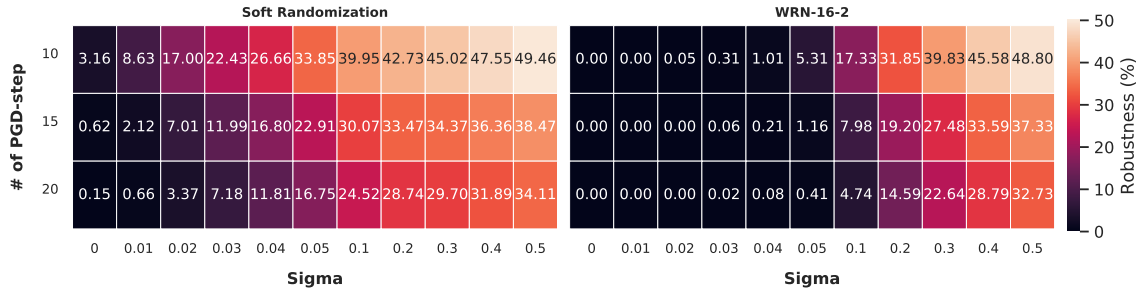

Figure 15: Soft-randomization (left) significantly increases the robustness of the student to PGD attack compared to both Hinton method (sigma=0) and WRN-16-2 trained with Gaussian data augmentation (right).

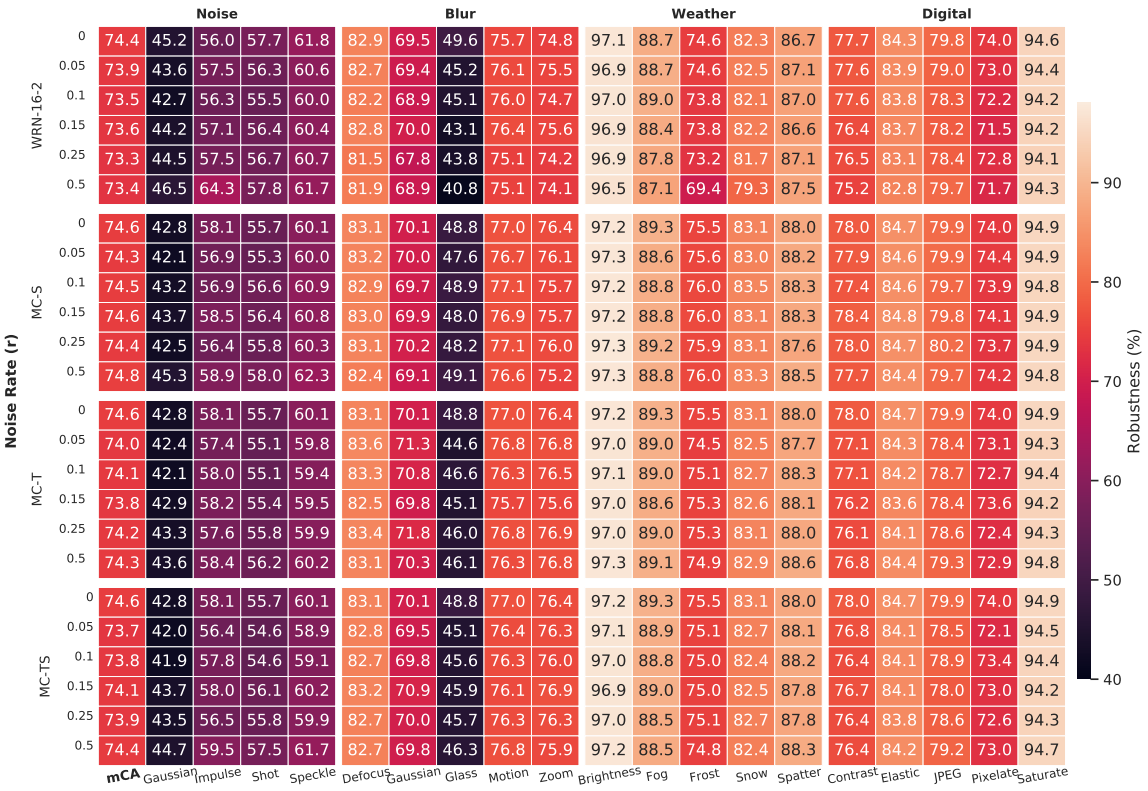

Figure 16: All variants of messy-collaboration achieve better robustness to natural perturbations compared to WRN-16-2 trained with Gaussian data augmentation. Furthermore, even at higher corruption levels, messy-collaboration maintains the natural robustness of Hinton method.

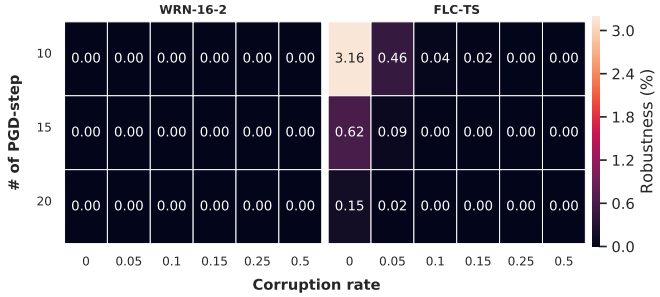

Figure 17: Fixed label corruption with corrupted teacher (FLC_TS) impairs the robustness to PGD attack.

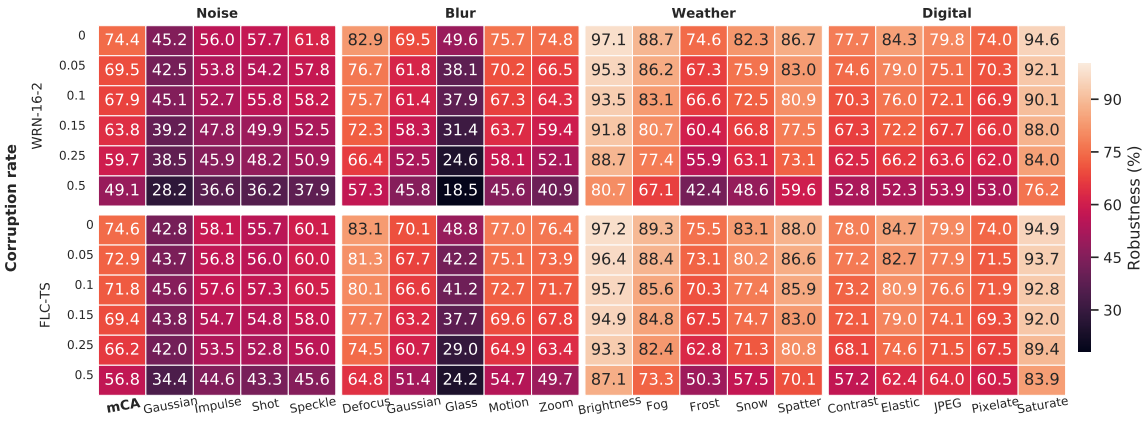

Figure 18: Fixed label corruption with corrupted teacher (FLC_TS) impairs the robustness to natural distortions.