
An Empirical Evaluation on Robustness and Uncertainty of Regularization Methods

Sanghyuk Chun¹ Seong Joon Oh² Sangdoon Yun¹ Dongyoon Han¹ Junsuk Choe³ Youngjoon Yoo¹

Abstract

Despite apparent human-level performances of deep neural networks (DNN), they behave fundamentally differently from humans. They easily change predictions when small corruptions such as blur and noise are applied on the input (lack of robustness), and they often produce confident predictions on out-of-distribution samples (improper uncertainty measure). While a number of researches have aimed to address those issues, proposed solutions are typically expensive and complicated (e.g. Bayesian inference and adversarial training). Meanwhile, many simple and cheap regularization methods have been developed to enhance the generalization of classifiers. Such regularization methods have largely been overlooked as baselines for addressing the robustness and uncertainty issues, as they are not specifically designed for that. In this paper, we provide extensive empirical evaluations on the robustness and uncertainty estimates of image classifiers (CIFAR-100 and ImageNet) trained with state-of-the-art regularization methods. Furthermore, experimental results show that certain regularization methods can serve as strong baseline methods for robustness and uncertainty estimation of DNNs.

1. Introduction

Recent studies have shown that inner mechanisms of DNNs are different from those of humans. For example, DNNs are easily fooled by human-imperceptible adversarial perturbations (adversarial robustness (Szegedy et al., 2013; Goodfellow et al., 2015)) and semantics-preserving transformations like noising, blurring, and texture corruptions (natural robustness (Geirhos et al., 2018b; Hendrycks & Dietterich, 2019; Geirhos et al., 2018a)). Another limitation of DNNs

is their inability to produce sound uncertainty estimates for their predictions. They are known to be inept at producing well-calibrated predictive uncertainties (known unknowns) and detecting out-of-distribution (OOD) samples (unknown unknowns) (Hendrycks & Gimpel, 2017).

For adversarial robustness, it has been shown that augmenting adversarial perturbations during training, or adversarial training, makes a model more adversarially robust (Kurakin et al., 2016; Madry et al., 2017; Xie et al., 2018). However, it is computationally challenging to employ it on large-scale datasets (Kurakin et al., 2016; Xie et al., 2018). Adversarially trained models overfit to the specific attack type used for training (Sharma & Chen, 2017), and the performance on unperturbed images drops (Tsipras et al., 2019). On the other hand, methods which improve robustness to non-adversarial corruptions are relatively less studied. Recently it is shown that training models by augmenting a specific noise enhances the performance on the target noise but can not be generalized to the other unseen noise types (Geirhos et al., 2018b). ImageNet-C dataset (Hendrycks & Dietterich, 2019) is proposed to evaluate robustness to 15 corruption types including blur and noise while a network should not observe the distortions during the train time. The authors have shown that the natural robustness is improved via adversarial training (Kannan et al., 2018) and Stylized ImageNet augmentation (Geirhos et al., 2018a), but have not considered more common and simpler regularization types; we provide those baseline experiments in this paper.

Efforts to improve uncertainty estimates of DNNs have followed two distinguishable paths: improving calibration of predictive uncertainty and out-of-distribution (OOD) sample detection. On the predictive uncertainty side, variants of Bayesian neural networks (Gal & Ghahramani, 2016; Kendall & Gal, 2017) and ensemble methods (Lakshminarayanan et al., 2017) have mainly been proposed. These approaches, however, are expensive and often require modifications of training and inference stages. On the OOD detection front, methods including threshold-based binary classifiers (Hendrycks & Gimpel, 2017) and real or GAN-generated OOD sample augmentation (Lee et al., 2018) have brought about improvements in OOD detections. Above approaches have demonstrated sub-optimal performances in

¹Clova AI Research, NAVER Corp. ²Clova AI Research, LINE Plus Corp. ³Yonsei University. Correspondence to: Sanghyuk Chun <sanghyuk.c@navercorp.com>.

our experiments, even compared to simple baselines.

As an independent line of research, many regularization techniques have been proposed to improve the generalization of DNN classifiers. For example, Batch Normalization (BN) (Ioffe & Szegedy, 2015) and data augmentation strategies such as random crop and random flip (Krizhevsky et al., 2012; Szegedy et al., 2016a) have become standard design choices for deep models. Despite their simplicity and efficiency, the effects of state-of-the-art regularization techniques such as label smoothing (Szegedy et al., 2016b), MixUp (Zhang et al., 2017) and CutMix (Yun et al., 2019) on the robustness and the uncertainty of deep models are still rarely investigated. A few works have shown indeed the effects of a few regularization techniques on DNN robustness (Zhang et al., 2017; Kannan et al., 2018; Yun et al., 2019), but we provide a more extensive analysis with both robustness and uncertainty perspectives.

We empirically evaluate state-of-the-art regularization techniques and show that they improve the classification, robustness, and uncertainty estimates for large-scale classifiers at marginal additional costs. We argue that certain regularization techniques must be considered as strong baselines for future researches in robustness and uncertainty of DNNs.

2. Revisiting Regularization Methods

In this section, we revisit several regularization methods including the state-of-the-art regularization methods used in our experiments.

Input augmentation: With proper data augmentation methods, a model can generalize better to the unseen samples. For example, random cropping and flipping are widely used to improve classification performances (Krizhevsky et al., 2012; Szegedy et al., 2016a; Huang et al., 2017). However, it is not always straightforward to distinguish augmentation types that improves the generalizability. For example, adversarial samples, geometric transformations, and pixel inversion are rarely helpful for improving classification performances (Tsipras et al., 2019; Cubuk et al., 2018). One of the most effective augmentation methods is Mixup (Zhang et al., 2017) which generates the in-between class samples by the pixel level interpolation. Another example of data augmentation is Cutout which erases pixels in a region sampled at random (DeVries & Taylor, 2017; Zhong et al., 2017). Recently proposed CutMix fills the pixels from other images instead of erasing pixels (Yun et al., 2019). While being simple and efficient, Mixup, Cutout and CutMix have shown significant improvements in classification performance. We consider their contribution to robustness and uncertainty estimates in our experiments.

Label perturbation: Deep models often suffer from overconfident predictions; they often produce predictions with

high confidence even on random Gaussian noise input (Hendrycks & Gimpel, 2017). One straightforward way to mitigate the issue is to penalize over-confident predictions by perturbing the target y . For example, label smoothing (Szegedy et al., 2016b) changes ground-truth label to a smoothed distribution whose probability of non-targeted labels are α/K , where α is a smoothing parameter whose default value is often 0.1 and K is the number of classes. By smoothing target predictions, models learn to regularize overconfident predictions. Another examples are Mixup (Zhang et al., 2017) and CutMix (Yun et al., 2019) which blend two one-hot labels into one smooth label by the mix ratio. Label smoothing is also known to offer a modest amount of robustness to adversarial perturbations (Kannan et al., 2018). It is thus widely used in adversarial training to achieve better adversarial robustness. We consider label smoothing as one of the axes for our investigation.

Other strategies for deep networks: Many researches have achieved more stable convergence and better generalization performance via weight regularization (weight decay) or feature-level manipulations like dropout (Srivastava et al., 2014) and Batch Normalization (Ioffe & Szegedy, 2015). Recently, randomly adding noises on intermediate features (Ghiasi et al., 2018; Gastaldi, 2017; Huang et al., 2016; Yamada et al., 2018), or adding extra paths to the model (Hu et al., 2017; 2018) have been proposed. We present robustness and uncertainty experiments on a selection of above regularization techniques.

3. Benchmarks for Robustness and Uncertainty Estimation

In this section, we describe the settings for the benchmarks used in our experiments. We tested four benchmarks: robustness to adversarial attacks, robustness to natural corruptions, robustness to occlusions, confidence calibration error, and out-of-distribution detection.

To evaluate adversarial robustness, we use FGSM (Goodfellow et al., 2015) with $\epsilon = 8/255$. Note that our baseline regularization methods cannot provide a provable defense to the adversarial attacks while adversarial training and ALP could mitigate the effect of the adversarial attacks.

For evaluating robustness against natural corruptions, we employ naturally corrupted ImageNet (ImageNet-C) proposed by (Hendrycks & Dietterich, 2019). In ImageNet-C, there are 15 transforms categorized into “noise”, “blur”, “weather”, and “digital” with five severities. For CIFAR-100 experiments, we create corrupted CIFAR-100 (CIFAR-100-C) using 75 transforms proposed in ImageNet-C. We report the average accuracy over all 75 transforms.

In occlusion robustness benchmarks, we generate occluded

Table 1. CIFAR-100 classification, robustness to adversarial and non-adversarial noises, and uncertainty benchmark results. For non-adversarial corruptions, we report top-1 error in CIFAR-100-C dataset and top-1 error in occluded CIFAR-100 test samples. We report out-of-distribution (OOD) detection errors averaged over seven OOD datasets. We fix the base architecture as PyramidNet-200 with $\alpha = 240$. LS stands for label smoothing. Lower is better for all reported numbers and all values are percentage.

Method	LS	Classification	Robustness			Uncertainty	
		CIFAR-100 Top-1 Err.	FGSM Top-1 Err.	CIFAR-C Top-1 Err.	Occlusion Top-1 Err.	Expected Calibration Err.	OOD Detection Err.
Baseline	–	16.45	84.20	45.11	72.19	8.00	18.05
	✓	16.73	82.82	46.50	74.40	2.51	17.59
ShakeDrop	–	15.08	77.91	44.37	78.69	8.01	19.76
	✓	15.05	63.09	43.74	82.22	2.53	25.59
Cutout	–	16.53	91.07	51.65	27.00	7.67	28.73
	✓	15.61	77.77	48.74	27.03	4.24	17.92
Cutout + ShakeDrop	–	15.91	88.66	50.00	26.19	6.63	19.55
	✓	13.49	69.59	43.86	26.33	1.45	18.40
Mixup	–	15.63	63.85	42.81	56.80	7.89	39.09
	✓	15.91	55.84	42.20	57.60	15.20	28.56
Mixup + ShakeDrop	–	14.91	61.91	40.60	57.07	7.28	22.92
	✓	14.79	56.32	40.32	56.76	15.85	18.54
CutMix	–	14.23	88.88	49.83	32.16	4.92	10.95
	✓	15.55	74.00	51.01	35.68	7.91	13.56
CutMix + ShakeDrop	–	13.81	70.75	43.36	35.83	2.46	19.82
	✓	13.83	62.72	44.99	34.96	5.26	18.89
Adversarial Logit Pairing	✓	24.75	51.32	50.04	92.27	6.67	21.57
Adversarial Training	✓	26.85	51.80	51.85	93.59	8.71	28.06
w/o Random Crop & Flip	–	21.83	90.63	48.71	77.46	7.99	26.91
Add Gaussian Noise	–	19.49	85.08	42.01	73.23	9.79	25.16
OOD augment (SVHN)	–	38.80	97.35	67.03	79.13	46.37	43.53
OOD augment (GAN)	–	34.78	94.65	57.09	85.30	38.22	33.35

samples by filling zeros (black pixels) over a square at the image center whose side length is half the image width; i.e., 16 for CIFAR-100 and 112 for ImageNet.

To show how the methods affect the confidence of predictions, we evaluate the expected calibration error (Guo et al., 2017). We view a classification system as a probabilistic confidence estimator whose confidence is a measurement of the trustworthy estimation. The bin size is set to 20. We refer (Guo et al., 2017) for further details of the evaluation.

Finally, we have tested the baseline OOD detection performance of each model. We have used the threshold-based detector proposed in (Hendrycks & Gimpel, 2017). Seven datasets used in (Liang et al., 2018) were considered: cropped Tiny ImageNet, resized Tiny ImageNet, cropped LSUN (Yu et al., 2015), resized LSUN, iSUN, Gaussian noise, and Uniform noise. We report the average detection error over the seven datasets.

4. Main Results

4.1. Training Settings

We first describe the settings for training models used in the robustness and the uncertainty benchmarks. To ensure the

effectiveness of each regularization methods, we employ a powerful baseline, PyramidNet-200 (Han et al., 2017) and ResNet-50 (He et al., 2016) for CIFAR-100 and ImageNet experiments, respectively.

We consider the state-of-the-art regularization methods of Cutout (DeVries & Taylor, 2017), Mixup (Zhang et al., 2017), CutMix (Yun et al., 2019), label smoothing (Szegedy et al., 2016b), ShakeDrop (Yamada et al., 2018), and their combinations for experiments. We optimize the models with the SGD with momentum. We set the batch size to 64 and training epochs to 300. The learning rate is initially set to 0.25 and is decayed by the factor of 1/10 at 150th and 225th epochs. We also employ random crop and random flip augmentations for all methods, unless specified otherwise.

For the comparison methods for adversarial robustness, we train the baseline model with adversarial training (Kurakin et al., 2016; Madry et al., 2017) and adversarial logit pairing (ALP) (Kannan et al., 2018). We use Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) with $\epsilon = 8/255$ as the threat model. All the results are evaluated with applying label smoothing to achieve better performances. We mix the clean and adversarial samples with the same ratio as proposed in (Kurakin et al., 2016). The optimizer for

Table 2. Comparison of noise augmentations on robustness to various noises. Noise, blur, weather and digital are a subset of CIFAR-C.

Methods	CIFAR-100 Top-1 Err.	FGSM Top-1 Err.	Occlusion Top-1 Err.	CIFAR-C mCE	Noise Top-1 Err.	Blur Top-1 Err.	Weather Top-1 Err.	Digital Top-1 Err.
Baseline	16.45	84.20	72.19	45.11	74.62	46.77	30.66	38.65
Adversarial Logit Pairing	24.75	51.32	92.27	50.04	69.94	51.75	40.62	44.70
Cutout	16.53	91.07	27.00	51.65	89.77	51.40	34.24	43.20
Add Gaussian Noise	19.49	85.08	73.23	42.01	54.63	48.42	31.54	38.48

Table 3. Comparison of well-regularized networks and baseline methods to improve robustness and uncertainty. SD stands for ShakeDrop.

Method	CIFAR-100 Top-1 Err.	FGSM Top-1 Err.	CIFAR-C Top-1 Err.	Occlusion Top-1 Err.	Expected Calibration Err.	OOD Detection Err.
Baseline	16.45	84.20	45.11	72.19	8.00	18.05
Cutout + SD + LS	13.49	69.59	43.86	26.33	1.45	18.40
Mixup + SD + LS	14.79	56.32	40.32	56.76	15.85	18.54
CutMix + SD + LS	13.83	62.72	44.99	34.96	5.26	18.89
Adversarial Logit Pairing	24.75	51.32	50.04	92.27	6.67	21.57
Add Gaussian Noise	19.49	85.08	42.01	73.23	9.79	25.16
OOD augment (SVHN)	38.80	97.35	67.03	79.13	46.37	43.53
OOD augment (GAN)	34.78	94.65	57.09	85.30	38.22	33.35

adversarial training is ADAM (Kingma & Ba, 2014).

As the baseline method for CIFAR-C, we consider Gaussian noise augmentation; the same type of perturbation taken from the CIFAR-C dataset (Hendrycks & Dietterich, 2019). For out-of-distribution (OOD) detection baseline, we augment OOD samples and the target labels to be the uniform label as proposed in (Lee et al., 2018). We augment two types of OOD samples used in (Lee et al., 2018): Street View House Numbers (SVHN) dataset and generated samples by GAN. In our experiments, we use WGAN-GP (Gulrajani et al., 2017) instead of DC-GAN (Radford et al., 2015).

All the experiments are done with NAVER Smart Machine Learning (NSML) (Sung et al., 2017; Kim et al., 2018).

4.2. CIFAR-100 Results

In this section, we evaluate the effects of the state-of-the-art regularization techniques on the various robustness and uncertainty benchmarks on CIFAR-100. We show that well-regularized models are powerful baselines.

In Table 1, we report classification, adversarial and natural robustness, and uncertainty measure evaluations. Classification performances are measured on CIFAR-100 test set; adversarial robustness is measured against the FGSM (Goodfellow et al., 2015) attack on CIFAR-100; natural robustness is measured on CIFAR-C (Hendrycks & Dietterich, 2019). Uncertainty qualities are measured in terms of expected calibration error (Guo et al., 2017) and OOD detection error rates (Hendrycks & Gimpel, 2017). We report the OOD detection errors at method-specific optimal thresholds.

Here we analyze the following questions from Table 1.

Can data augmentation improve robustness against various

perturbations at once? Data augmentation is a straightforward solution to improve robustness against specific type of noise, e.g., adversarial perturbation, Gaussian noise, and occlusion. In Table 2, we have observed that different type of augmentation methods improve robustness against the target noise. For example, ALP improves adversarial robustness but it fails to improve robustness against occlusion and other natural corruptions. Similarly, in Table 2, Cutout is only method that improves occlusion robustness among the other augmentation methods. However, Cutout degrades other types of robustness, such as adversarial robustness, compare to the baseline. By adding Gaussian noise to the input, robustness to the common corruptions is enhanced, especially for the “noise”. In summary, we have observed that it would be difficult to improve the robustness against various type of corruptions at once. A similar phenomenon was also observed by (Geirhos et al., 2018b).

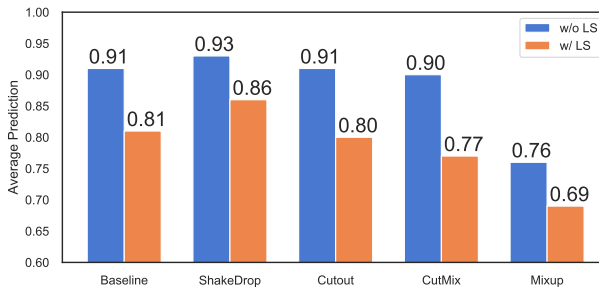


Figure 1. Average top-1 prediction probability by models trained with state-of-the-art regularization methods. Models with label smoothing (LS) produce less confident predictions.

Can label smoothing help adversarial robustness and uncertainty estimates? In our experiments, adding label smoothing (LS) alone does not generally improve classification

Table 4. Top-1 errors of considered regularization techniques on various test-time perturbations. We report the average Top-1 error among clean images, FGSM attacked images, occluded images, and naturally corrupted images (ImageNet-C). Finally, we report mCE (mean corruption error) normalized by AlexNet. SD and LS stand for ShakeDrop and label smoothing, respectively.

	Average	Clean	FGSM	Occ.	Noise	Blur	Weather	Digital	mCE
Baseline	67.43	23.68	91.85	46.01	78.58	86.63	64.99	80.24	77.55
Label Smoothing	62.67	22.31	73.60	44.35	77.08	82.30	61.72	77.33	74.44
ShakeDrop	64.57	22.03	87.19	42.98	76.13	83.42	61.56	78.69	74.87
ShakeDrop + LS	61.45	21.92	72.65	42.85	74.47	82.15	60.47	75.67	73.10
Cutout	64.81	22.93	88.50	29.72	79.94	85.37	65.34	81.87	78.01
Cutout + LS	61.90	22.02	75.24	29.08	79.80	84.51	62.72	79.93	76.54
Mixup	61.46	22.58	75.60	44.20	73.09	81.49	58.83	74.42	71.88
Mixup + LS	58.54	22.41	69.43	42.31	65.36	82.95	53.37	73.94	69.14
CutMix	62.08	21.60	69.04	30.09	80.88	84.87	64.11	83.95	78.29
CutMix + LS	61.02	21.87	67.41	31.51	77.01	84.61	63.13	81.56	76.55
CutMix + SD	61.75	21.60	80.00	31.28	77.06	84.18	61.04	77.07	74.69
CutMix + SD + LS	60.96	21.90	68.65	31.62	76.04	84.53	62.82	81.16	76.14

accuracies. Surprisingly, however, we observe that LS improves robustness against adversarial perturbation, calibration error, and OOD detection performance (Table 1). For example, by adding LS, Cutout + ShakeDrop achieves 13.49% classification top-1 error and FGSM top-1 error 69.59% where performances without LS are 15.91% and 88.66% for classification and adversarial robustness respectively. We believe that it is because a model trained with LS produces low confident predictions in general (Figure 1). In particular, LS shows impressive improvements in the expected calibration error, except for Mixup and CutMix families. We believe the result is due to the fact that Mixup and CutMix already contain the label mixing stage that already lowers the prediction confidences; further adding label smoothing makes the overall confidences too low.

Can well-regularized models be a powerful baseline for the robustness and uncertainty estimations? In Table 3, we have observed that our well-regularized models such as Cutout + ShakeDrop + label smoothing, Mixup + ShakeDrop + label smoothing, and CutMix + ShakeDrop outperform methods targeted for improved robustness and uncertainty estimations (ALP and OOD augmentations) in many evaluation metrics. For example, ALP model shows occlusion top-1 error 92.27% while Cutout and CutMix based models show 26.33% and 34.96% top-1 error respectively. It is notable that OOD augmentations are not effective for CIFAR-100 tasks, while they have been shown to be effective for toy datasets like SVHN and CIFAR-10 (Lee et al., 2018).

4.3. ImageNet Experiments

In this section, we report experimental results on ImageNet. We use ResNet-50 (He et al., 2016) as the baseline model and train the models with same training scheme as used in (Yun et al., 2019). We only evaluate robustness bench-

marks, i.e., adversarial robustness against FGSM, natural robustness against ImageNet-C, and robustness to occlusion.

In Table 4, we report the top-1 error on clean images, attacked images, occluded images, naturally corrupted images (subsets of ImageNet-C), and their average. Also we report the mCE (mean corrupted error) normalized by AlexNet (Krizhevsky et al., 2012) which is proposed in (Hendrycks & Dietterich, 2019).

As we observed in CIFAR-100 experiments, regularized models provide better overall performances. For example, CutMix achieves 62.08% average error alone but adding ShakeDrop and LS improves average error to 60.96%. Table 4 also shows that label smoothing is still effective in improving the robustness of the models in ImageNet experiments. Mixup helps robustness against common corruptions; CutMix shows better classification performance, adversarial robustness, and occlusion robustness.

Interestingly, in our experiments, Mixup + label smoothing achieves the state-of-the-art performance on ImageNet-C mCE of 69.14% where current best model is stylized-ImageNet trained model (Geirhos et al., 2018a) with mCE of 69.3%. Note that stylized-ImageNet requires heavy pre-computations to generate the stylized images, and requires additional fine-tuning on ImageNet data.

Methods used in our experiments improve the overall robustness and uncertainty performances at negligible additional costs. We believe that well-regularized models should be considered as powerful baselines for the robustness and the uncertainty estimation benchmarks.

5. Conclusion

In this paper, we have empirically compared the robustness and uncertainty estimates of state-of-the-art regularization methods against prior methods specifically designed for such aspects. We have observed that methods proposed to solve the specific problem are only effective on their targeted task. For example, adversarial training only improves adversarial robustness while it degrades classification performance, robustness against common corruptions and occlusion, and uncertainty estimates. On the other hand, good combinations of simple and cheap regularization techniques improve overall robustness and uncertainty estimation performances, and even surpass specialized methods in certain uncertainty and robustness tasks. We believe that well-regularized models have largely been overlooked in robustness and uncertainty studies, and that they should be considered as powerful baselines in future works.

References

- Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.
- DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.
- Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.
- Gastaldi, X. Shake-shake regularization. In *arXiv:1705.07485*, 2017.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018a.
- Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 7538–7550, 2018b.
- Ghiasi, G., Lin, T.-Y., and Le, Q. V. Dropblock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 10750–10760, 2018.
- Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017.
- Han, D., Kim, J., and Kim, J. Deep pyramidal residual networks. In *CVPR*, 2017.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.
- Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=HJz6tiCqYm>.
- Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.
- Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *arXiv:1709.01507*, 2017.
- Hu, J., Shen, L., Albanie, S., Sun, G., and Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 9423–9433, 2018.
- Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Deep networks with stochastic depth. In *ECCV*, 2016.
- Huang, G., Liu, Z., and Weinberger, K. Q. Densely connected convolutional networks. In *CVPR*, 2017.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.
- Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5574–5584. Curran Associates, Inc., 2017.
- Kim, H., Kim, M., Seo, D., Kim, J., Park, H., Park, S., Jo, H., Kim, K., Yang, Y., Kim, Y., et al. NSML: Meet the mlaas platform with a real-world case study. *arXiv preprint arXiv:1810.09957*, 2018.

- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6402–6413. Curran Associates, Inc., 2017.
- Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=ryiAv2xAZ>.
- Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.
- Sharma, Y. and Chen, P.-Y. Attacking the madry defense model with l_1 -based adversarial examples. *arXiv preprint arXiv:1710.10733*, 2017.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.
- Sung, N., Kim, M., Jo, H., Yang, Y., Kim, J., Lausen, L., Kim, Y., Lee, G., Kwak, D., Ha, J.-W., et al. NSML: A machine learning platform that enables you to focus on your models. *arXiv preprint arXiv:1712.05902*, 2017.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- Szegedy, C., Ioffe, S., and Vanhoucke, V. Inception-v4, inception-resnet and the impact of residual connections on learning. In *ICLR Workshop*, 2016a.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016b.
- Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=SyxAb30cY7>.
- Xie, C., Wu, Y., van der Maaten, L., Yuille, A., and He, K. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018.
- Yamada, Y., Iwamura, M., Akiba, T., and Kise, K. Shake-drop regularization for deep residual learning. *arXiv preprint arXiv:1802.02375*, 2018.
- Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.
- Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. *arXiv preprint arXiv:1905.04899*, 2019.
- Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
- Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.