Self-Supervised Learning for Robust Multimodal Neural Networks

Mr.E.Sivarajan[1], Assistant Professor
Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Morur (Post), Sankari (Tk), Salem, Tamil Nadu, India
sivarajan.e@shanmugha.edu.in
Mr.V.Mouliraj[2], Assistant Professor
Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Morur (Post), Sankari (Tk), Salem, Tamil Nadu, India
mouliraj.v@shanmugha.edu.in
Mrs.M.Ramya[3], Assistant Professor
Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Morur (Post), Sankari (Tk), Salem, Tamil Nadu, India
ramya.m@shanmugha.edu.in
Mr.Saravanakumar Pichumani[4], Assistant Professor
Department of Computer Science and Engineering, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Morur (Post), Sankari (Tk), Salem, Tamil Nadu, India
saravanakumar@shanmugha.edu.in

Abstract- Self-supervision provides effective representations for downstream tasks without requiring labels. However, existing approaches lag behind fully supervised training and are often not thought beneficial beyond obviating or reducing the need for annotations. We find that self-supervision can benefit robustness in a variety of ways, including robustness to adversarial examples, label corruption, and common input corruptions. Additionally, self-supervision greatly benefits out-of-distribution detection on difficult, near-distribution outliers, so much so that it exceeds the performance of fully supervised methods. These results demonstrate the promise of self-supervision for improving robustness and uncertainty estimation and establish these tasks as new axes of evaluation for future self-supervised learning research.

Keywords – Self-supervision, representations for downstream tasks, requiring labels, for improving robustness and uncertainty estimation.

INTRODUCTION

Self-supervised learning holds great promise for improving representations when labeled data are scarce. In semi-supervised learning, recent self-supervision methods are state-of-the-art [Gidaris et al., 2018, Dosovitskiy et al., 2016, Zhai et al., 2019], and self-supervision is essential in video tasks where annotation is costly [Vondrick et al., 2016, 2018]. To date, however, self-supervised approaches lag behind fully supervised training on standard accuracy metrics and research has existed in a mode of catching up to supervised performance [1]. Additionally, when used in conjunction with fully supervised learning on a fully labeled dataset, self-supervision has little impact on accuracy [2]. This raises the question of whether large labeled datasets render self-

supervision needless. We show that while self-supervision does not substantially improve accuracy when used in tandem with standard training on fully labeled datasets, it can improve several aspects of model robustness, including robustness to adversarial examples [Madry et al., 2018], label corruptions [Patrini et al., 2017, Zhang and Sabuncu, 2018], and common input corruptions such as fog, snow, and blur [Hendrycks and Dietterich, 2019]. Importantly, these gains are masked if one looks at clean accuracy alone, for which performance stays constant. Moreover [3], we find that self-supervision greatly improves out-of-distribution detection for difficult, near-distribution examples, a long-standing and underexplored problem. In fact, using self-supervised learning techniques on CIFAR-10 and ImageNet for out-of-distribution detection, we are even able to surpass fully supervised methods. These results demonstrate that self-supervision need not be viewed as a collection of techniques allowing models to catch up to full supervision [4]. Rather, using the two in conjunction provides strong regularization that improves robustness and uncertainty estimation even if clean accuracy does not change. Importantly, these methods can improve robustness and uncertainty estimation without requiring larger models or additional data [Schmidt et al., 2018, Kurakin et al., 2017]. They can be used with task-specific methods for additive effect with no additional assumptions. With self-supervised learning, we make tangible progress on adversarial robustness, label corruption, common input corruptions, and out-of-distribution detection, suggesting that future self-supervised learning methods could also be judged by their utility for uncertainty estimates and model robustness [5].

RELATED WORK

(i) Self-supervised learning- A number of self- supervised methods have been proposed, each exploring a different pretext task. Doersch et al. [2015] predict the relative position of image patches and use the resulting representation to improve object detection. Dosovitskiy et al. [2016] create surrogate classes to train on by transforming seed image patches. Similarly, Gidaris et al. [2018] predict image rotations (Figure 1). Other approaches include using colorization as a proxy task [Larsson et al., 2016], deep clustering methods [Ji et al., 2018], and methods that maximize mutual information [Hjelm et al., 2019] with high-level representations [van den Oord et al., 2018, Hénaff et al., 2019]. These works focus on the utility of self-supervision for learning without labeled data and do not consider its effect on robustness and uncertainty[6].

(ii) Robustness- Improving model robustness refers to the goal of ensuring machine learning models are resistant across a variety of imperfect training and testing conditions. Hendrycks and Dietterich [2019] look at how models can handle common real-world image corruptions (such as fog, blur, and JPEG compression) and propose a comprehensive set of distortions to evaluate real-world robustness. Another robustness problem is learning in the presence of corrupted labels [Nettleton et al., 2010, Patrini et al., 2017]. To this end, Hendrycks et al. [2018] introduce Gold Loss Correction (GLC), a method that uses a small set of trusted labels to improve accuracy in this setting[7]. With high degrees of label corruption, models start to overfit the misinformation in the corrupted labels [Zhang and Sabuncu, 2018, Hendrycks et al., 2019a], suggesting a need for ways to supplement training with reliable signals from unsupervised objectives. Madry et al. [2018] explore adversarial robustness and propose PGD adversarial training, where models are trained with a minimax robust optimization objective. Zhang et al. [2019] improve upon this work with a modified loss function and develop a better understanding of the trade-off between adversarial accuracy and natural accuracy[8].

(iii) Out-of-distribution detection- Out-of-distribution detection has a long history. Traditional methods such as one-class SVMs [Schölkopf et al., 1999] have been revisited with deep

representations [Ruff et al., 2018], yielding improvements on complex data. A central line of recent exploration has been with out-of-distribution detectors using supervised representations. Hendrycks and Gimpel [2017] propose using the maximum softmax probability of a classifier for out-of-distribution detection [18]. Lee et al. [2018] expand on this by generating synthetic outliers and training the representations to flag these examples as outliers [9]. However, Hendrycks et al. [2019b] find that training against a large and diverse dataset of outliers enables far better out-of-distribution detection on unseen distributions. In these works, detection is most difficult for near-distribution outliers, which suggests a need for new methods that force the model to learn more about the structure of in-distribution examples.

ROBUSTNESS

(i) Robustness to Adversarial Perturbations- Improving robustness to adversarial inputs has proven difficult, with adversarial training providing the only longstanding gains [Carlini and Wagner, 2017, Athalye et al., 2018][17]. In this section, we demonstrate that auxiliary self-supervision in the form of predicting rotations [Gidaris et al., 2018] can improve upon standard Projected Gradient Descent (PGD) adversarial training [Madry et al., 2018][10]. We also observe that self-supervision can provide gains when combined with stronger defenses such as TRADES [Zhang et al., 2019] and is not broken by gradient-free attacks such as SPSA [Uesato et al., 2018].

	Clean	20-step PGD	100-step PGD
Normal Training	94.8	0.0	0.0
Adversarial Training	84.2	44.8	44.8
+ Auxiliary Rotations (Ours)	83.5	50.4	50.4

Table 1- Results for our defense[6]

(ii) Method- We explore improving representation robustness beyond standard PGD training with auxiliary rotation-based self-supervision in the style of Gidaris et al. [2018]. In our approach [16], we train a classification network along with a separate auxiliary head, which takes the penultimate vector from the network as input and outputs a 4-way softmax distribution. This head is trained along with the rest of the network to predict the amount of rotation applied to a given input image (from 0 °, 90°, 180°, and 270°). Our overall loss during training can be broken down into a supervised loss and a self-supervised loss [11].

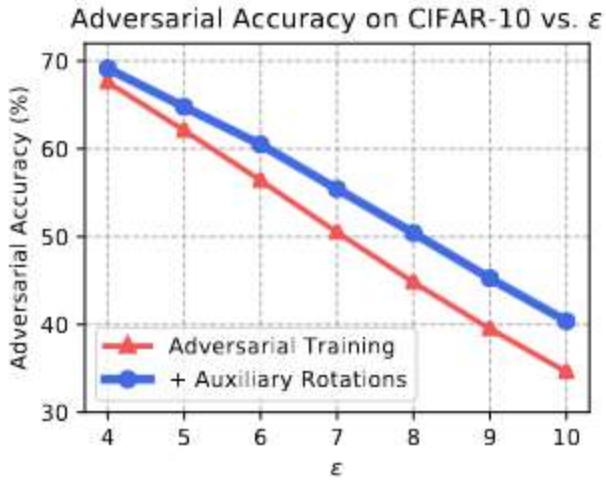

Figure 1- The effect of attack strength on a ε =8/255 adversarially trained model[8]

ROBUSTNESS TO LABEL CORRUPTIONS

(i) Setup- Training classifiers on corrupted labels can severely degrade performance. Thus, several prior works have explored training deep neural networks to be robust to label noise in the multi-class classification setting Sukhbaatar et al. [2014], Patrini et al [15-23]. [2017], Hendrycks et al. [2018]. We use the problem setting from these works. Let x, y, and ˜y be an input, clean label, and potentially corrupted label respectively. Given a dataset ˜D of (x, ˜y) pairs for training, the task is to obtain high classification accuracy on a test dataset Dtest of cleanly-labeled (x, y) pairs [12].

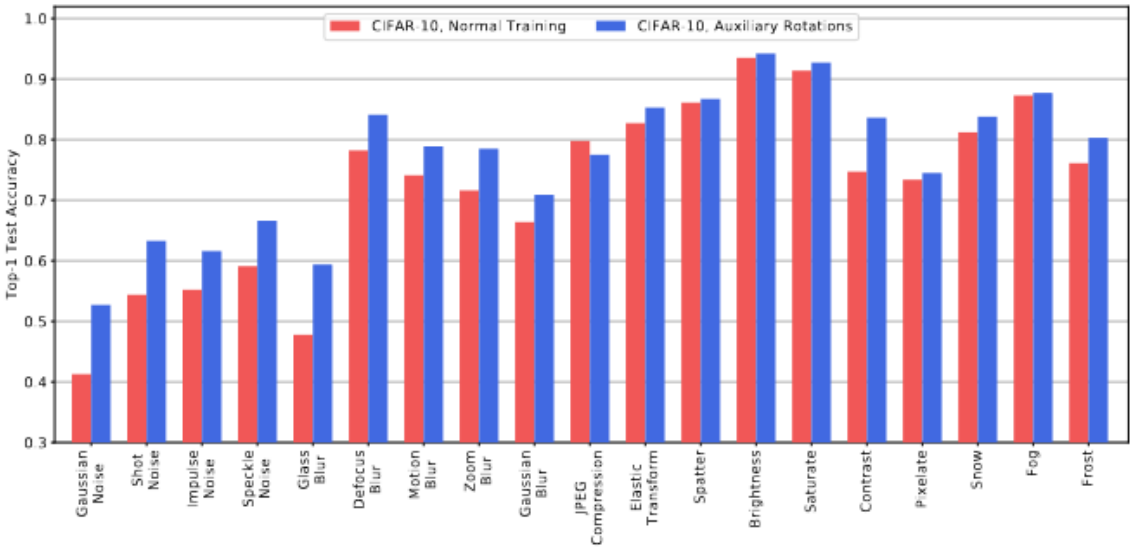

Figure 2- A comparison of the accuracy of usual training compared to training with auxiliary rotation self-supervision on the nineteen CIFAR-10-C corruptions[10]

(ii) Methods- Training without loss correction methods or self-supervision serves as our first baseline, which we call No Correction in Table 2. Next, we compare to the state-of-the-art Gold Loss Correction (GLC) Hendrycks et al. [2018]. This is a two-stage loss correction method based on Sukhbaatar et al. [2014] and Patrini et al. [2017]. The first stage of training estimates the matrix C of conditional corruption probabilities, which partially describes the corruption process. The

second stage uses the estimate of C to train a corrected classifier that performs well on the clean label distribution. The GLC assumes access to a small dataset of trusted data with cleanly-labeled examples. Thus, we specify the percent of amount of trusted data available in experiments as a fraction of the training set. This setup is also known as a semi-verified setting Charikar et al. [2017] [13-15].

(iii) Analysis- We observe large gains in robustness from auxiliary rotation prediction. Without loss corrections, we reduce the average error by 5.6% on CIFAR-10 and 5.2% on CIFAR-100. This corresponds to an 11% relative improvement over the baseline of normal training on CIFAR-100 and a 26% relative improvement on CIFAR-10. In fact, auxiliary rotation prediction with no loss correction outperforms the GLC with 5% trusted data on CIFAR-100. This is surprising given that the GLC was developed specifically to combat label noise [14].

	CIFAR-10		CIFAR-100	
	Normal Training	Rotations	Normal Training	Rotations
No Correction	27.4	21.8	52.6	47.4
GLC (5% Trusted)	14.6	10.5	48.3	43.2
GLC (10% Trusted)	11.6	9.6	39.1	36.8

Table 2- Label corruption results comparing normal training to training with auxiliary rotation self-supervision[12]

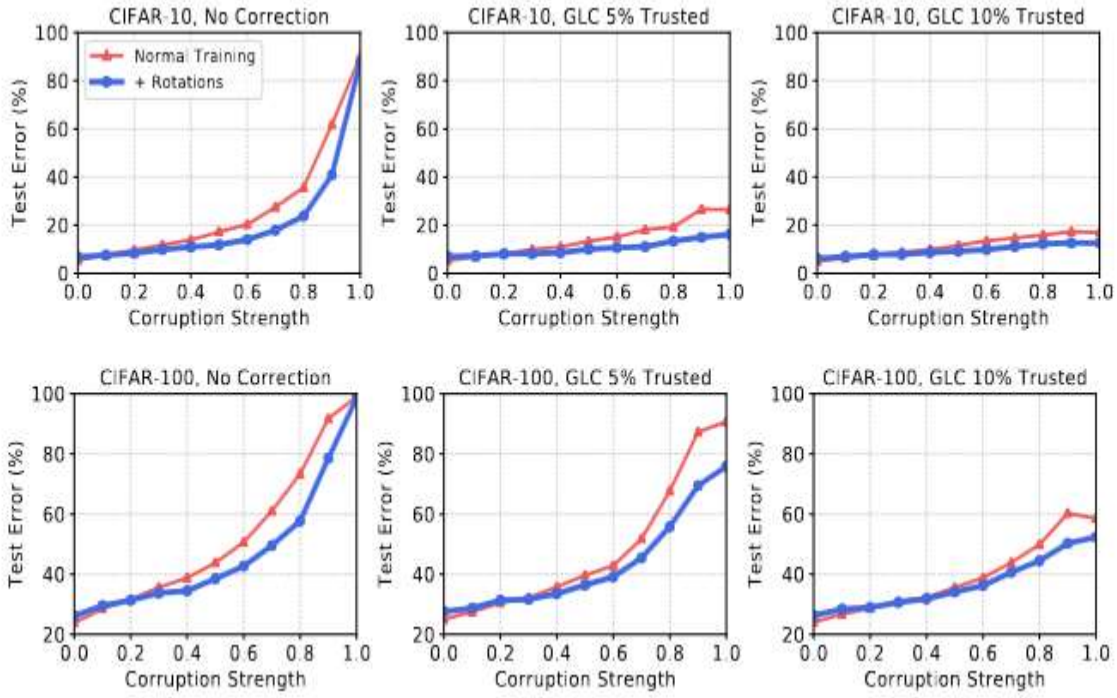

Figure 3- Error curves for label corruption comparing normal training to training with auxiliary rotation self-supervision[15]

CONCLUSION

In this paper, we applied self-supervised learning to improve the robustness and uncertainty of deep learning models beyond what was previously possible with purely supervised approaches. We found large improvements in robustness to adversarial examples, label corruption, and common input corruptions. For all types of robustness that we studied, we observed consistent gains by supplementing current supervised methods with an auxiliary rotation loss. We also found that self- supervised methods can drastically improve out-of-distribution detection on difficult, near-distribution anomalies, and that in CIFAR and ImageNet experiments, self-supervised methods outperform fully supervised methods. Self-supervision had the largest improvement over supervised techniques in our ImageNet experiments, where the larger input size meant that we were able to apply a more complex self-supervised objective. Our results suggest that future work in building more robust models and better data representations could benefit greatly from self-supervised approaches.

REFERENCES

[1]. Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Proceedings of the 35th International Conference on Machine Learning, ICML 2023, July 2023.

[2]. Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. ArXiv, abs/1811.00995, 2024.

[3]. Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods, 2023.

[4]. Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. STOC, 2024.

[5]. Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In International Conference on Machine Learning, 2023.

[6]. Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. CVPR, 2024.

[7]. Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision, pages 1422–1430, 2023.

[8]. Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox.Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE transactions on pattern analysis and machine intelligence, 38(9):1734–1747, 2024.

[9]. Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In International Conference on Learning Representations, 2024.

[10]. Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. CoRR, abs/1805.10917, 2023.

[11]. Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. ICLR, 2023.

[12]. Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. ICLR, 2024.

[13]. Sanjay Kumar Suman, Dhananjay Kumar and L. Bhagyalakshmi, “Non Cooperative Power Control Game with New Pricing for Wireless Ad hoc Networks”, International Review on Computers and Software, vol. 9, no. 1, pp. 18-28, 2014. ISSN: 1828-6003,

[14]. S. Porselvi, Sanjay Kumar Suman and L. Bhagyalakshmi, "Harvesting RF energy for mobile charging", Australian Journal of Basic and Applied Science, vol. 9, no. 20, pp. 454-465, June 2015.

[15]. K. Swapna, P. Rajalakshmi and Sanjay Kumar Suman, "Security Enhancement in MANET using Game Theory", Middle East Journal of Scientific Research, vol. 23, pp. 190-195, 2015.

[16]. VinaySrivatsan, Sanjay Kumar Suman, L. Bhagyalakshmi and S. Porselvi, "Non radiative wireless power transfer", Journal of Advances in Natural and Applied Sciences, vol. 10, no. 16, pp. 147-153, Nov. 2016.

[17]. Sujeetha Devi, Bhagyalakshmi L and Sanjay Kumar Suman, "Cluster based energy efficient joint routing algorithm for delay minimization in wireless sensor networks", International Journal of Pure and Applied Mathematics, vol. 119, no. 15, 307-313, 2018

[18]. Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. NeurIPS, 2023.

[19]. Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. Proceedings of the International Conference on Machine Learning, 2023a.

[20]. Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In International Conference on Learning Representations, 2023b.

[21]. R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In International Conference on Learning Representations, 2024.

[22]. Olivier J. Hénaff, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding, 2023.

[23]. Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information distillation for unsupervised image segmentation and clustering. CoRR, abs/1807.06653, 2023.