SUMMER SCHOOL 2023

Self Supervised Learning for Robustness and OOD Detection

Nick Würflein

nick.wuerflein@uni-jena.de

ai.uni-jena.de

Introduction

Neural Network (NN) Models can be used to solve various tasks, including image classification and object segmentation. Usually, these tasks are solved using a so-called supervised learning approach, in which training data is passed through the network and the model output is compared to previously known labels. Another approach is to use so-called self-supervised learning (SSL). In this scenario, no labels are previously known; rather, a label is automatically generated during the learning process.

The goal of out-of-distribution (OOD) detection is to detect if new input data is drawn from the probability distribution modeled by the NN. This can be useful in the deployment of NN models in critical situations, for example in autonomous driving or in the medical field, because inputs that are not modeled by the NN can be refused and errors can be avoided. Another aspect that can be improved by using SSL methods is model robustness. Broadly speaking, model robustness describes how well a model can work around different types of noise.

In this contribution, we want to highlight SSL and its advantages for increased robustness in image tasks, as well as improved OOD detection compared to the respective state-of-the-art methods. To this end, we summarize the work of Chen et al. as well as Hendrycks et al.

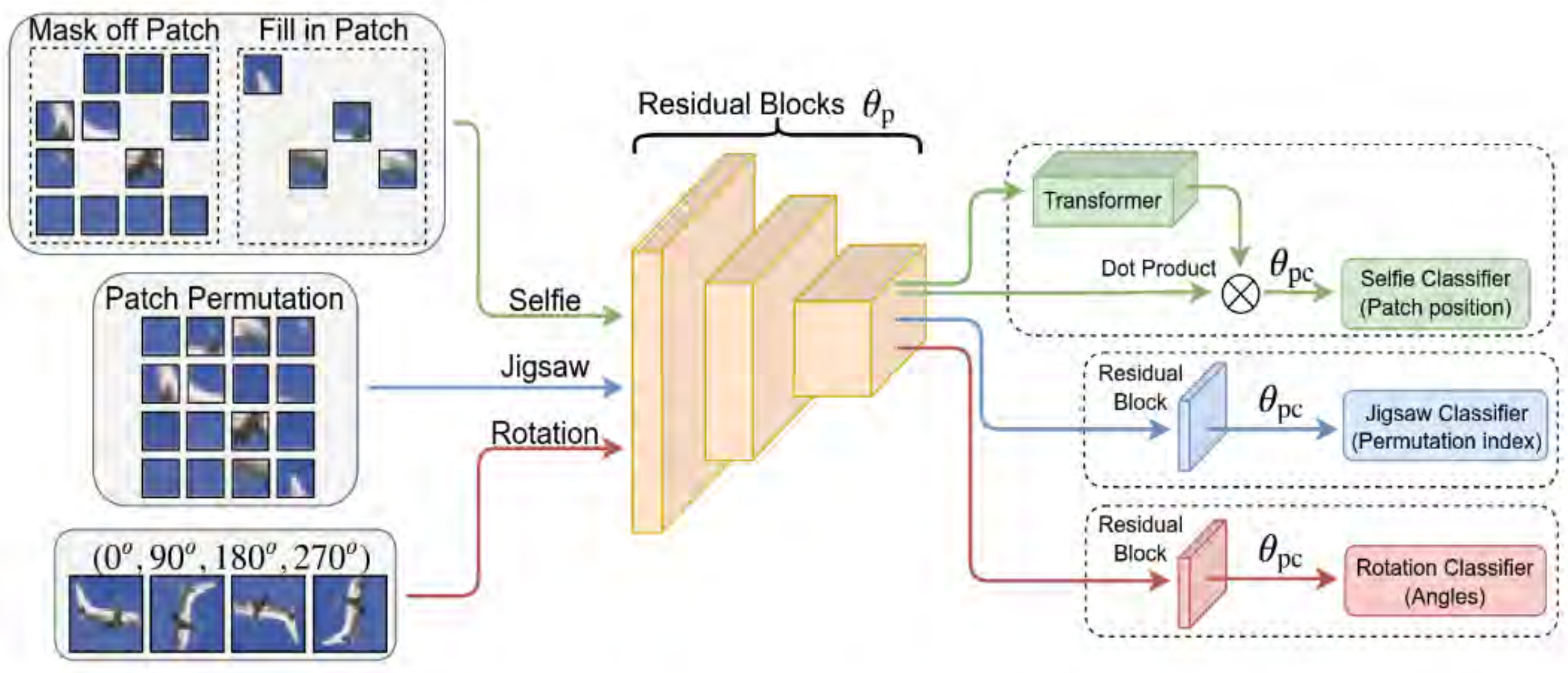
Self Supervised Learning:

By solving the SSL task, the model is expected to learn the underlying structures in the pictures. For example, to detect the rotation of an object, it is necessary for the model to learn to localize objects, recognize their orientation, and differentiate different types of objects. As stated earlier, SSL doesn't require a-priori labels since labels are generated automatically. This is done by alternating the data in a previously defined way, for example, by rotating the picture or by dividing the picture into different patches and rearranging them randomly. Afterward the model is tasked with predicting the alteration, i.e., classifying the rotation or predicting the permutation indices.

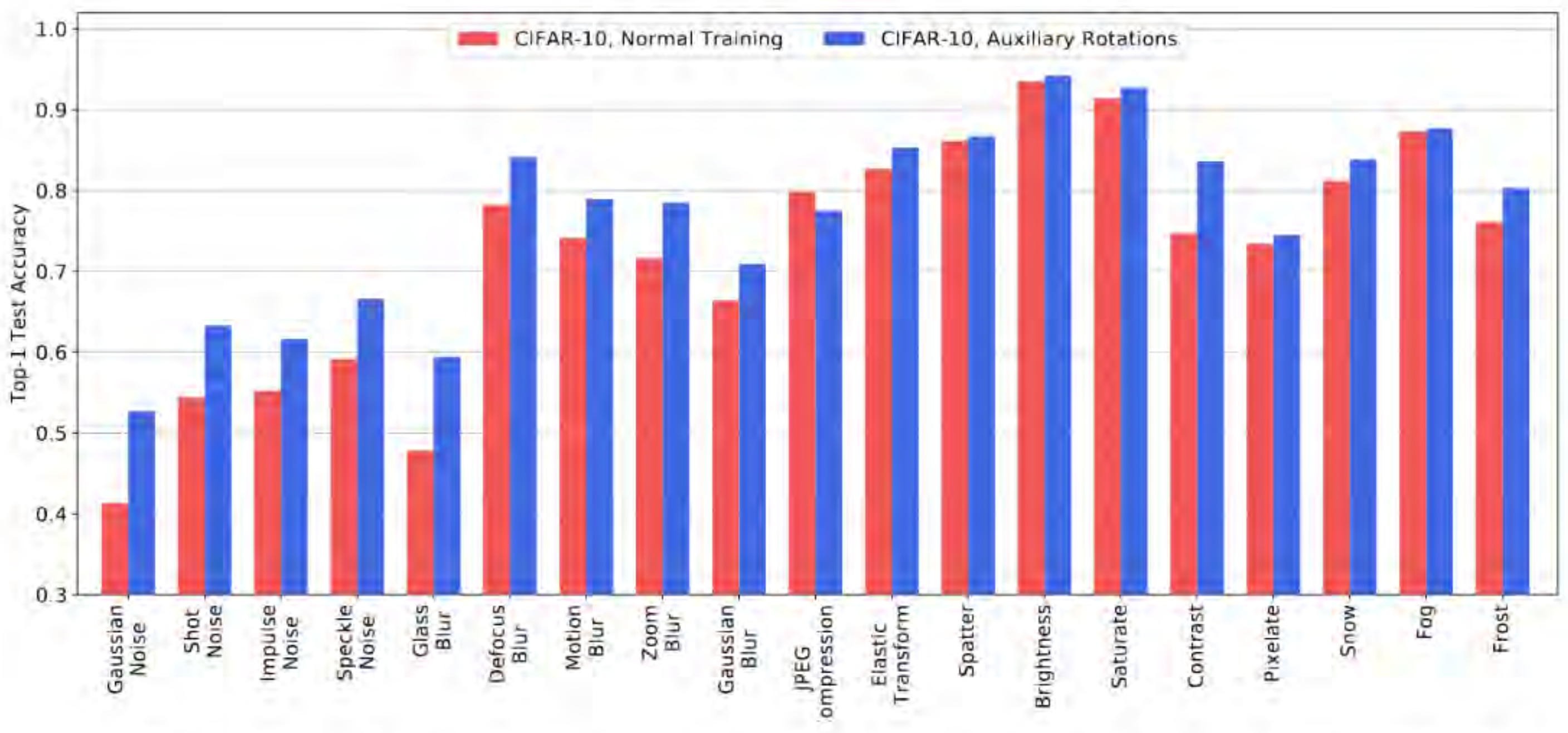
We look at two different kinds of SSL setups:

1. Hybrid: A SL setup is augmented by an additional classification head, which classifies the alteration. An additional error term from this classification is added to the SL error.

2. Pre-training: A NN model is trained on the SSL classification task via self-supervised pre-training. After the SLL training is done, the output of layers prior to the SLL classification is used as features for the true task, e.g., image classification or segmentation. Alternatively, the pre-trained model is fine-tuned to the downstream task.



Different SSL techniques used for pre-training. Selfie masks parts of the input and learns to fill in the empty spots. Jigsaw scrambles the data and learns to recreate the original image. Rotation rotates the picture and learns the orientation.



Top-1 Test Accuracy for a series of common corruptions on the CIFAR-10 data set. Normal training and hybrid SSL using image rotations are compared.

Out of Distribution Detection:

For OOD detection, a one-class detector as well as a mutliclass OOD detector are described. To keep it simple, we limit the description to one class detector. To detect out-of-distribution examples, first, k SSL classifiers are trained for all k classes using the image rotation approach. During this training, only one class is used per classifier. In the inference step, a new input is rotated in every direction, and the sum of all probabilities of the correct rotation are calculated. This sum, the so-called anomaly score, is then used in a second classification step to predict if a given input is in or out of distribution. The anomaly score is higher for OOD examples compared to in-distribution examples.

Compared to the state-of-the-art method for training OOD classifiers, so-called outlier exposure, the described SSL approach achieves a higher AUROC (area under receiver operating characteristic) value.

Method	AUROC
Supervised (OE)	56.1
RotNet	65.3
RotNet + Translation	77.9
RotNet + Self-Attention	81.6
RotNet + Translation + Self-Attention	84.8
RotNet + Translation + Self-Attention + Resize (Ours)	85.7

Results for one-class OOD detection averaged over all classes as AUROC (area under receiver operating characteristic) on a 30 class subset of ImageNet. The supervised model is trained with OE (Outlier Exposure). Translation and resize are additional SSL techniques.

Model Robustness:

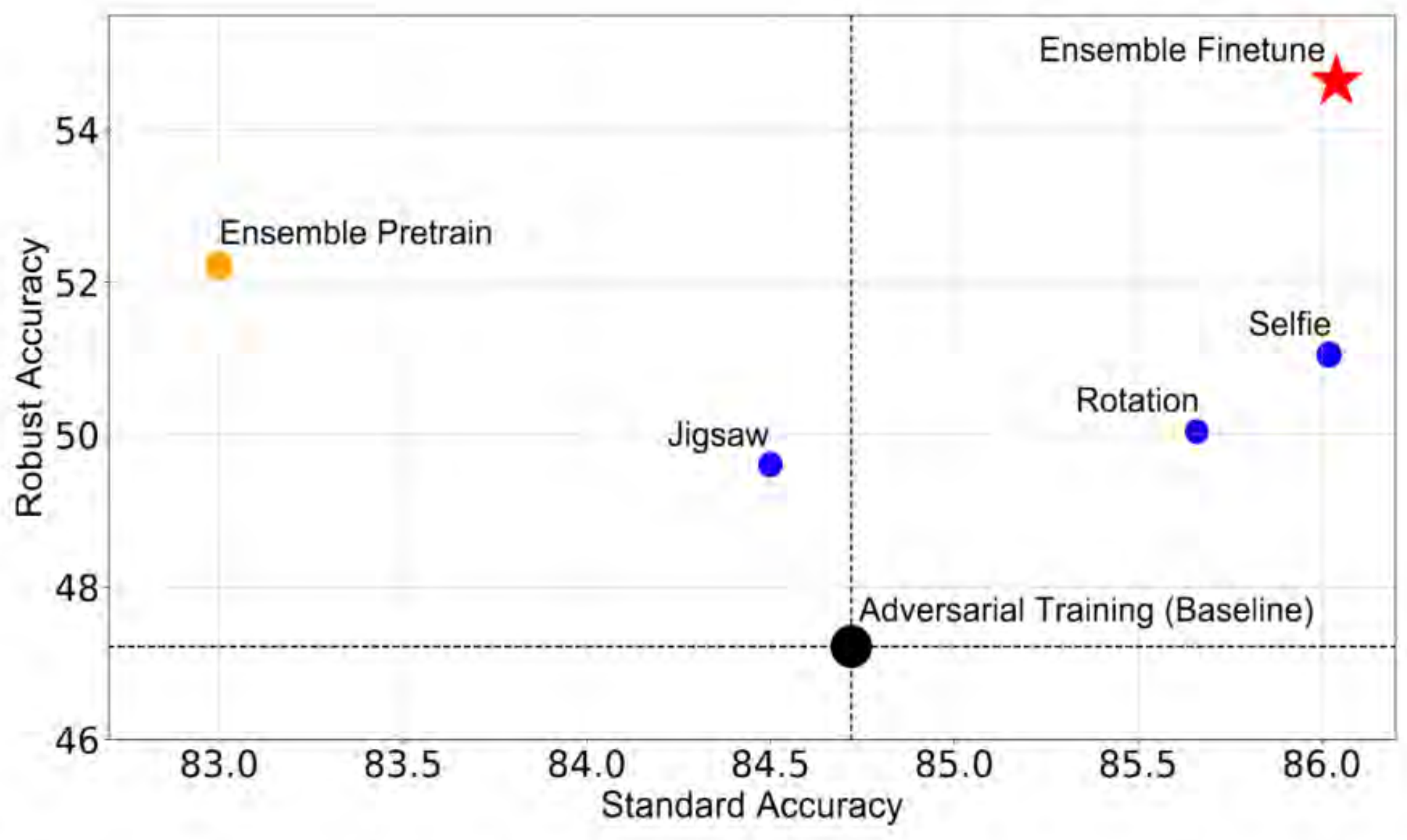
Robustness can be achieved against different types of imperfect training or testing conditions, like:

1. Label Noise: The dataset has errors. Some data is labeled incorrectly.

2. Corruptions: Different types of noise. Gaussian noise, blur, fog, altering brightness or contrast, etc.

3. Adversarial Examples: Specific attacks on the model that maximize error with minimal changes in the input.

Using either a hybrid or a pre-training approach increases the robustness against adversarial examples, even compared to the state-of-the-art method. Improvements in robustness, as demonstrated by the increase in robust accuracy, are made without the additional cost of reducing clean accuracy.



Robust and standard accuracy of the different types of model configurations. The Ensemble Pretrain model is not fine-tuned. Jigsaw, Rotation and Selfie are fine-tuned, but only use one type of SSL respectively. Ensemble Finetune combines all SSL techniques and uses fine-tuning afterward.

Discussion and Conclusion

There are several advantages to using SSL approaches, both in hybrid and pre-training form.

- Compared to outlier exposure for OOD detection, no big outlier dataset is needed, and the training is usually less computationally expensive.
- Compared to gold loss correction for increasing robustness against label noise, no additional trusted dataset is needed.
- Robustness against corruptions is increased for nearly all types of corruptions.
- Improvements seem to stack over different types of additional techniques for improving robustness or OOD detection. If additional improvements for OOD and robustness are needed, other techniques may improve the results even further.
- Drawbacks to applying the SSL methods are:
- So far, these positive effects have only been observed in image data. Implementing SSL techniques in other types of data, e.g., video or 3D sensory data, may not be as successful, and there is no standard way of doing that.
- Modifications are needed to accommodate SSL in tasks that are already solved by SL approaches.

References:

- Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. CoRR, abs/2003.12862, 2020

- Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. CoRR, abs/1906.12340, 2019.