**Benchmarking Neural Network Robustness to Common Corruptions and Perturbations**

1 INTRODUCTION

2 RELATED WORK

**Adversarial Examples** An adversarial image is a clean image perturbed by a small distortion carefully crafted to confuse a classifier.

**Robustness in Speech**

**ConvNet Fragility Studies** Several studies demonstrate the fragility of convolutional networks on simple corruptions.

**Robustness Enhancements**

3 CORRUPTIONS, PERTURBATIONS, AND ADVERSARIAL PERTURBATIONS

We now define corruption and perturbation robustness and distinguish them from adversarial perturbation robustness. Corruption robustness measures the classifier’s average-case performance on corruptions C, while adversarial robustness measures the worst-case performance on small, additive, classifier-tailored perturbations.

4 THE IMAGENET-C AND IMAGENET-P ROBUSTNESS BENCHMARKS

4.1 THE DATA OF IMAGENET-C AND IMAGENET-P

**IMAGENET-C Design** The corruptions are drawn from four main categories— noise, blur, weather, and digital. Each corruption type has five levels of severity since corruptions can manifest themselves at varying intensities.

**Common Corruptions**

**IMAGENET-P Design** Like IMAGENET-C, IMAGENETPconsists of noise, blur, weather, and digital distortions. IMAGENET-P departs from IMAGENET-C by having perturbation sequences generated from each ImageNet validation image. Each sequence contains more than 30 frames, so we counteract an increase in dataset size and evaluation time by using only 10 common perturbations.

**Common Perturbations**

4.2 IMAGENET-C AND IMAGENET-P METRICS AND SETUP

**IMAGENET-C Metrics**

**IMAGENET-P Metrics**

**Preserving Metric Validity** The goal of IMAGENET-C and IMAGENET-P is to evaluate the robustness of machine learning algorithms on novel corruptions and perturbations.

5 EXPERIMENTS

5.1 ARCHITECTURE ROBUSTNESS

Yet this is not to suggest that there is a fundamental trade-off between corruption and perturbation robustness. In fact, both corruption and perturbation robustness can improve together, as we shall see later.

5.2 ROBUSTNESS ENHANCEMENTS

**Histogram Equalization**

**Multiscale Networks** Multiscale architectures achieve greater corruption robustness by propagating features across scales at each layer rather than slowly gaining a global representation of the input as in typical convolutional neural networks.

**Feature Aggregating and Larger Networks** Some recent models enhance the ResNet architecture by increasing what is called feature aggregation.

**Stylized ImageNet**

**Adversarial Logit Pairing**

6 CONCLUSION

In this work, we found several methods to increase robustness, introduced novel experiments and metrics, and created new datasets for the rigorous study of model robustness, a pressing necessity as models are unleashed into safety-critical real-world settings.