Poster: WED-AM-390

Seasoning Model Soups for Robustness to Adversarial and Natural Distribution Shifts

Francesco Croce*, Sylvestre-Alvise Rebuffi, Evan Shelhamer, Sven Gowal

U. Tübingen DeepMind

*intern at DeepMind

Overview

Problems

Robustness to particular Lp-bounded attacks does not generalize to other attacks

Adversarially trained models are not robust to other distribution shifts

Adapting the type of robustness requires retraining

Our solution

Step 1. Start with a single Lp-robust model

Step 2. Fine-tune it to different threat models

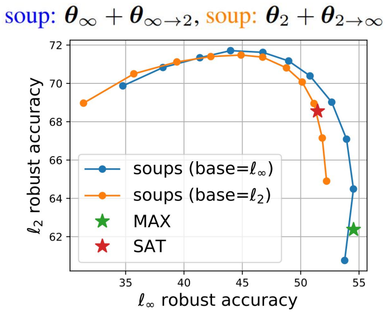
Step 3. Make the soup!

Using linear combinations of model parameters:

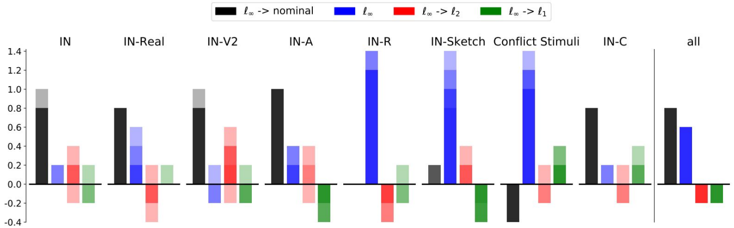
$$\theta_{\text{soup}} = w \cdot \theta_p + (1 - w) \cdot \theta_q$$

Results

Control level and type of robustness without retraining



Find a soup for each distribution shift



Motivation

Problem: deep networks are vulnerable to *adversarial attacks*, small input perturbations that result in errors

Solution: adversarial training [Madry et al., 2018] gives robust models against Lp-bounded attacks

But…

- Robustness to specific Lp-bounded attacks **does not generalize** to other attacks
- Adversarially trained models are **not robust to natural distribution shifts**
- Adapting type of robustness needs **retraining**

Idea: a short fine-tuning (even 1 epoch) of an Linf-robust model can give classifiers robust w.r.t. L2 or L1 threats or high clean performance [Croce & Hein, 2022]

… is it possible to efficiently combine these various models?

Soups of Lp-robust models

Classifiers fine-tuned from a single robust model to other threat models can be merged via **linear combination** of the parameters

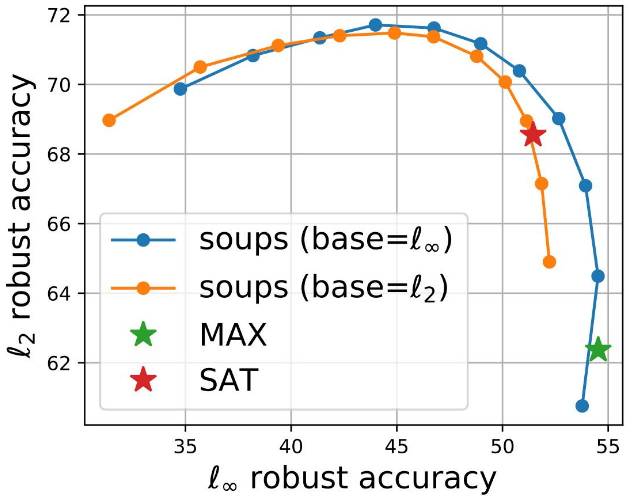
This enables **soups** [Wortsmann et al., 2022] of models with different types of robustness

$$\theta_{\text{soup}} = w \cdot \theta_p + (1 - w) \cdot \theta_q$$

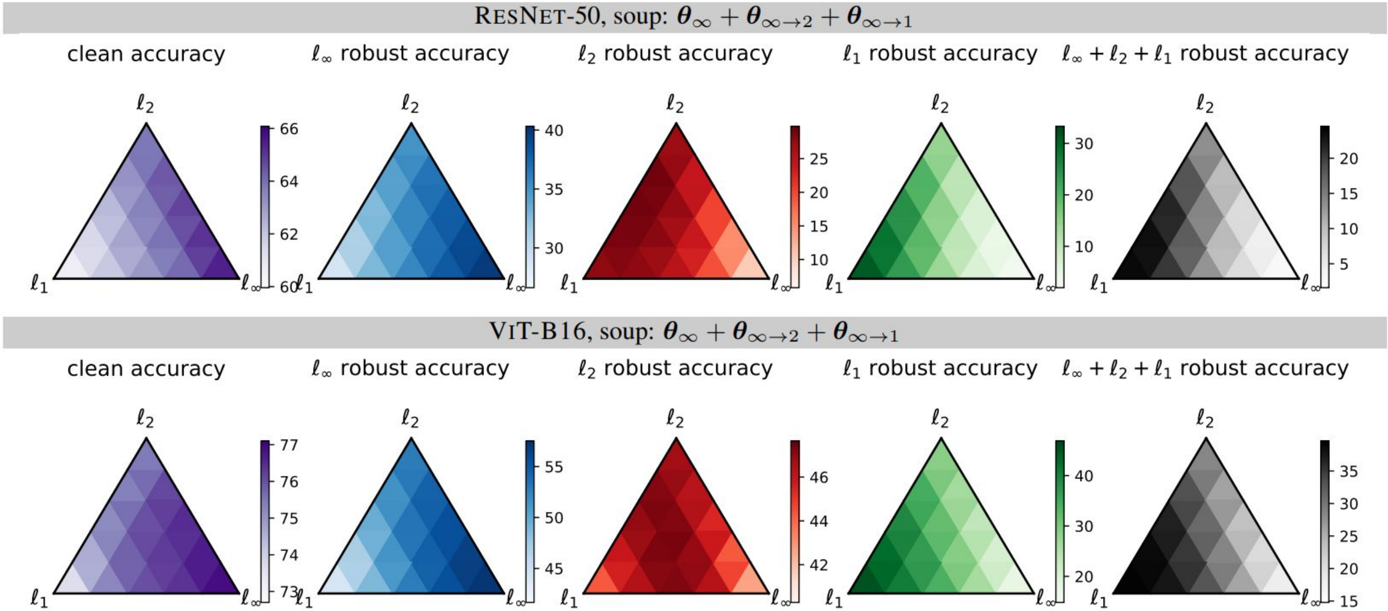
We can **control the trade-off** between types of robustness via the interpolation weight **without training** additional models!

Example. Soups of Linf and L2 robust models (CIFAR-10, WideResNet-28-10).

soup: $\boldsymbol{\theta}_\infty + \boldsymbol{\theta}_{\infty\to2}$, soup: $\boldsymbol{\theta}_2 + \boldsymbol{\theta}_{2\to\infty}$



We can do the same for **three threat models**, robust w.r.t. Linf, L2 and L1 (for various architectures and datasets, e.g. ImageNet below)

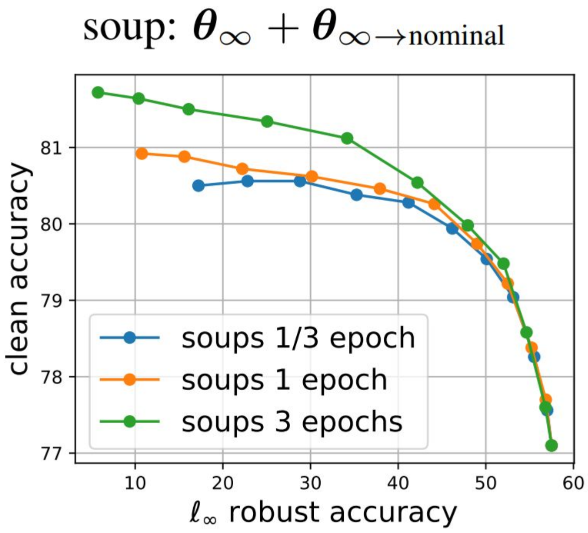


Soups of nominal and robust models

We can also make model soups of nominal and robust models to balance clean performance and robustness.

Longer fine-tuning improves the front formed by the soups

Example. Soups of nominal and Linf robust models (ImageNet, ViT-B).

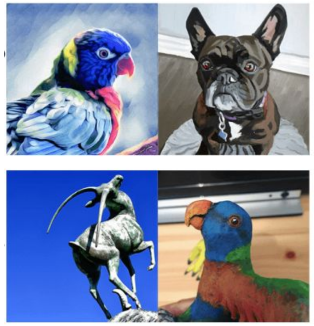


Soups for distribution shifts

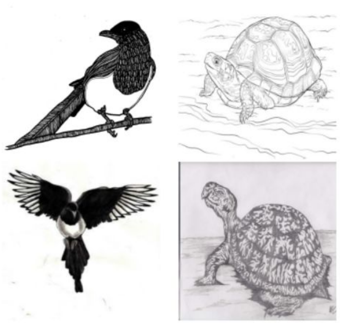
Problem: the performance of classifiers might deteriorate in the presence of shifts like ImageNet-R or ImageNet-Sketch

Goal: we want to find a soup which performs well on the new distribution

ImageNet-R



ImageNet-Sketch



Our framework:

1. Four base models

- Linf robust
- Linf → L2
- Linf → L1
- Linf → nominal

2. Soup selection

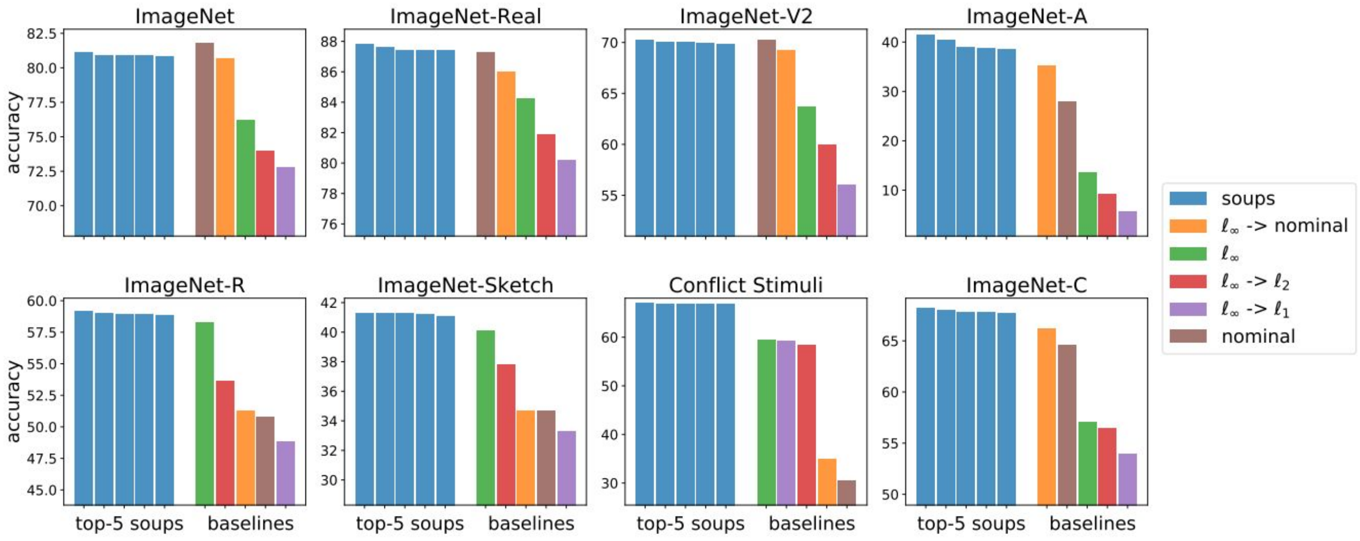
- collect a few labelled images with the shift
- select best interpolation weights (grid search)

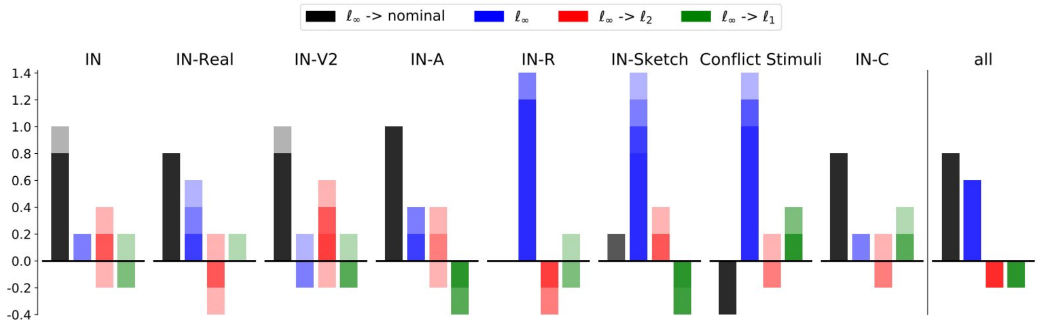
3. Test on unseen images

Test the model soup selected on the adaptation set on unseen validation images

Soups for distribution shifts: results

- We test 8 datasets (ImageNet and various shifts)
- The best individual model varies across datasets
- In most cases the soups outperform the base models
- The soups composition changes according to the dataset





the composition of the soups varies across datasets

a single soup for all datasets

SETUP	# FP	IMAGENET	IN-REAL	IN-V2	IN-A	IN-R	IN-SKETCH	CONFLICT STIMULI	IN-C	MEAN
Baselines										
Nominal training	×1	82.64%	87.33%	71.42%	28.03%	47.94%	34.43%	30.47%	64.45%	55.84%
Adversarial training	×1	76.88%	83.91%	64.81%	12.35%	55.76%	40.11%	59.45%	55.44%	56.09%
Fine-tuned MAE-B16	×1	83.10%	88.02%	72.80%	37.92%	49.30%	35.69%	27.81%	63.23%	57.23%
AdvProp	×1	83.39%	88.06%	73.17%	34.81%	53.04%	39.25%	38.98%	70.39%	60.14%
Pyramid-AT	×1	83.14%	87.82%	72.53%	32.72%	51.78%	38.60%	37.27%	67.01%	58.86%
Indep. networks ensemble	×2	82.86%	87.78%	71.73%	25.99%	54.20%	37.33%	46.41%	65.61%	58.99%
Individual networks ensemble	×4	81.31%	86.97%	70.21%	23.13%	54.82%	39.51%	56.02%	68.17%	60.02%
Fixed grid search on 1000 images										
Single soup	×1	82.49%	87.85%	71.99%	34.31%	53.84%	39.84%	38.52%	66.82%	59.46%
Dataset-specific soups	×1	82.29%	87.89%	71.95%	38.27%	56.39%	40.73%	67.03%	69.34%	(64.24%)

a different soup for each dataset