Interpreting Global Perturbation Robustness

Róisín Luo[†,‡] James McDermott[†,‡] Colm O'Riordan[†,‡]

[‡] University of Galway [†] National Centre for Research Training in AI (CRT-AI, Ireland)

Introduction

Research question: Mechanistically interpreting why some vision models are more robust to perturbations than others.

Key takeaways:

- **Feature signal-to-noise (SNR) Bias**: (1) Robust features (RFs) $\Longleftrightarrow$ low-frequency signals (LFs); (2) Non-Robust Features (NRFs) $\Longleftrightarrow$ high-frequency signals (HFs).
- **Feature Robustness Modulus**: Features can be categorized into robust features (RFs) or non-robust features (NRFs) by high SNRs or low SNRs on spectra.
- **Role of Feature SNRs**: Models trained on higher SNR features tend to have higher robustness.
- **Mechanistic interpretability**: Measuring the frequency response of a model can explain its robustness.

Feature SNRs have a spectral bias

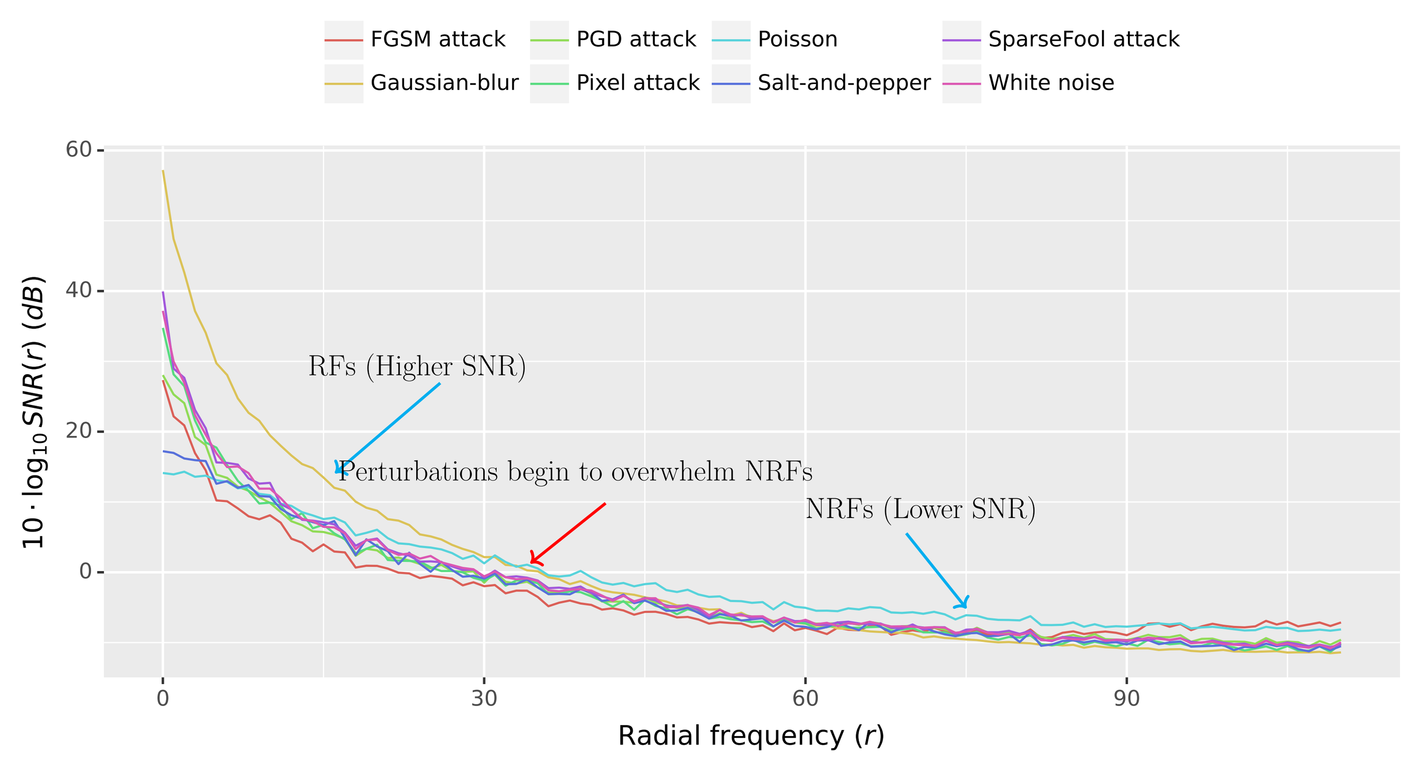

Figure: Feature SNR Bias.

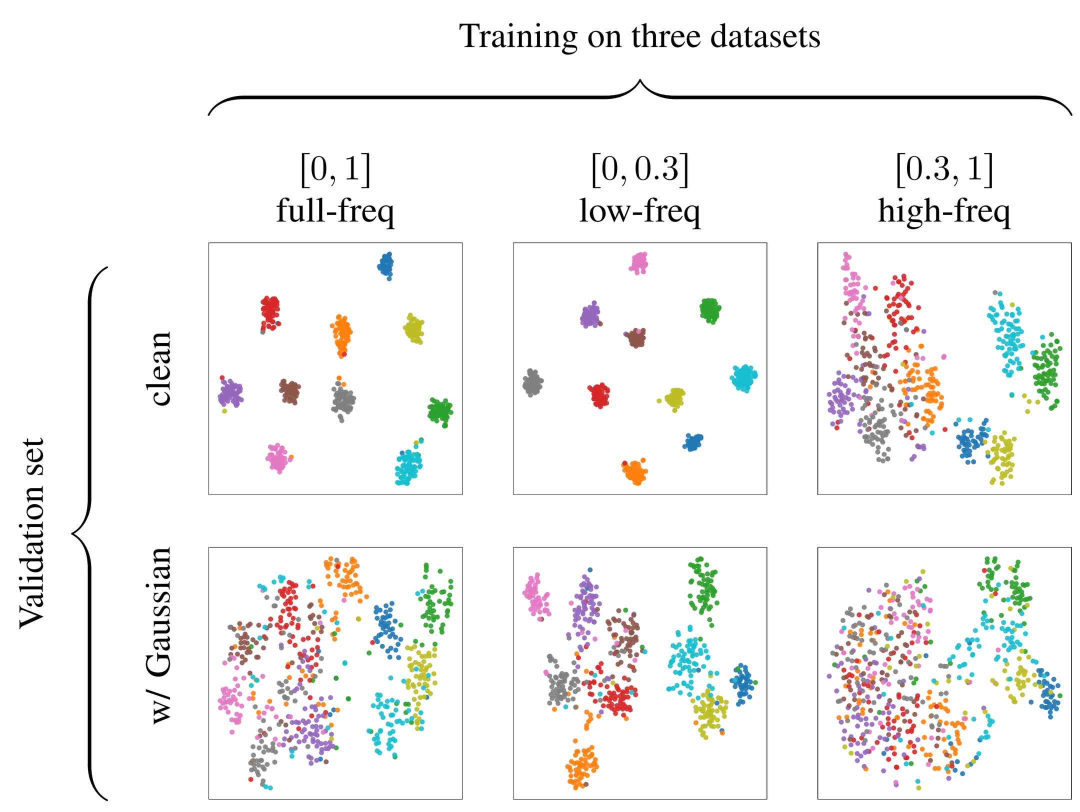

Figure: Role of Feature SNRs in training.

Global feature marginal information contribution on spectra

Mutual information

Frequency index

Network

Baseline

$$\text{Information contribution of } I_i \longrightarrow \phi_f(I_i) = \int_{I \setminus I_i} \{\mathbb{I}\,[f(X_{S \cup I_i}); Y] - \mathbb{I}\,[f(X_S); Y]\}\, \mathrm{d}\mathbb{P}(S)$$

Signal w/o I_i

Label

Probability measure on $S \in I \setminus I_i$

Experiments and implications

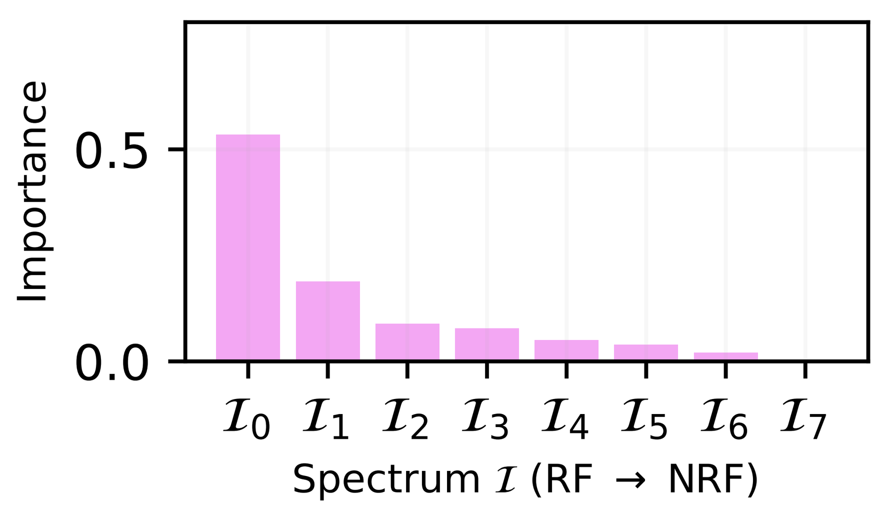

Figure: Frequency response (resnet18).

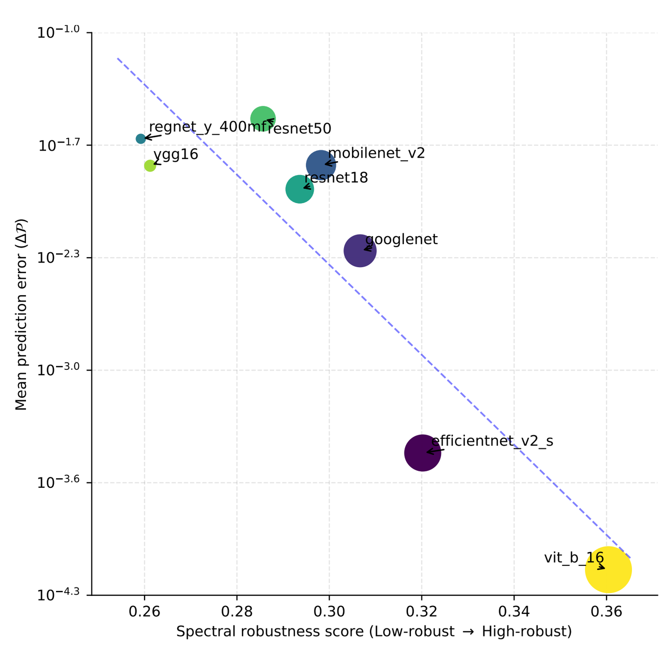

Figure: Scalarized $\{\phi_f(I_i)\}_i$ correlates with adversarial perturbation.

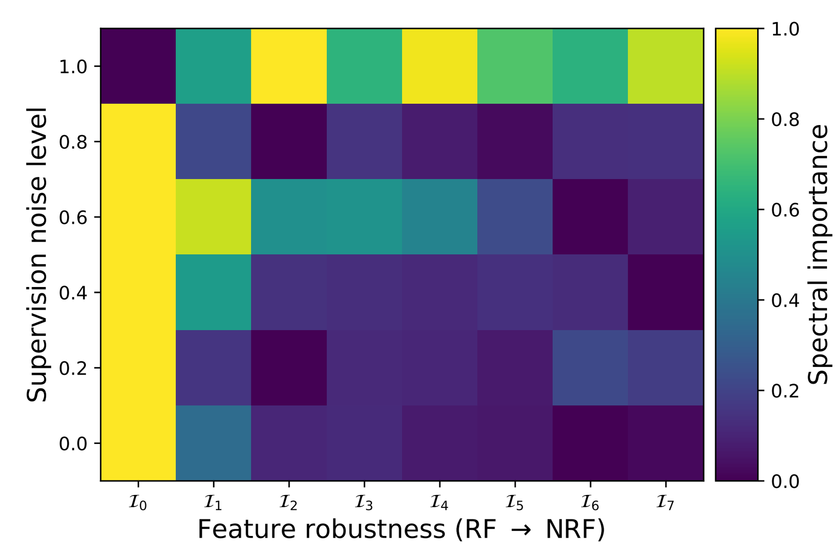

Figure: Model responses to label noise during training.

Acknowledgment

This publication has emanated from research conducted with the financial support of **Taighde Éireann** - Research Ireland under Grant number 18/CRT/6223.

Paper LinkedIn

HOST INSTITUTION

PARTNER INSTITUTIONS