The Effect of Intrinsic Dataset Properties on Generalization: Unraveling Learning Differences Between Natural and Medical Images

Nicholas Konz, Maciej A. Mazurowski

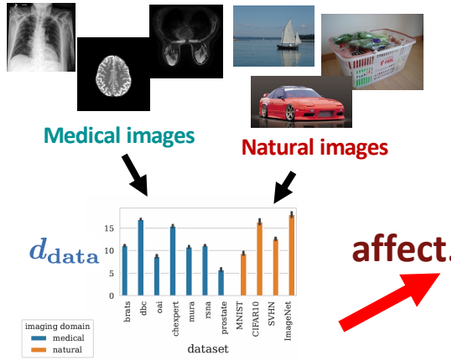
Duke University

ICLR

Q: **How and *why* do neural networks learn differently from medical vs. natural images?**

A: **Measurable intrinsic properties of the training set affect generalization behavior of the trained model!**

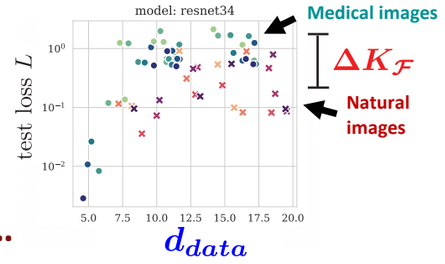
Intrinsic properties of the training set



affect...

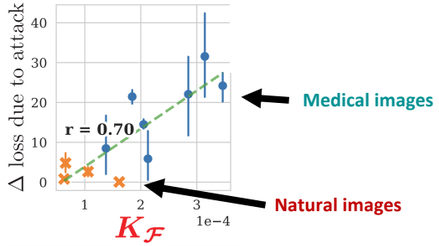
Generalization ability

$$\log L \lesssim -\frac{1}{d_{data}} \log N + \log K_{\mathcal{F}} + a$$



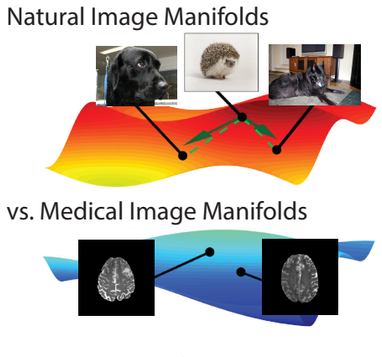
Adversarial robustness

$$\log \hat{R}(f) \gtrsim -\log K_{\mathcal{F}} + \log b$$



Background

- Generalization error of a trained network scales with the intrinsic manifold dim. of the training set.
- But the ***steepness*** of this curve empirically differs for medical vs. natural image models! ***Why??***
- Similar discrepancies found for adversarial attack susceptibility...
- Our goal: understand and model this from a **scientific perspective.**



Modified from Buchanan et al., ICLR 21

Main Contributions

- Establish **generalization and adversarial robustness scaling laws** (left) **that depend on measurable intrinsic dataset properties**: intrinsic dimension and ***label sharpness*****,** a metric we introduce**:**

$$\hat{K}_{\mathcal{F}} := \max_{j,k} \left(\frac{|y_j - y_k|}{||x_j - x_k||} \right)$$

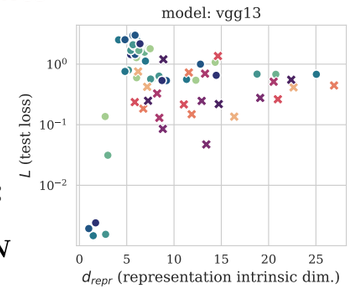
label sharpness: *Describes how similar images can be while still being from different classes.*

- We empirically validate the scaling laws on **6** models, **11** datasets, and **7** training set sizes.
- Our results show that **medical images** typically have much higher label sharpness, leading to the generalization discrepancy. This also makes them **more susceptible to adversarial attack!**
- Overall, we provide and validate the first theoretical models for the gap in deep learning behavior between natural and medical images.

Additional Results

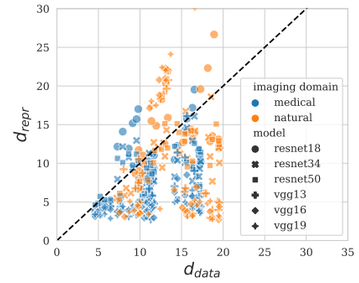
Generalization scaling law with respect to ***learned representation*** **intrinsic dimension:**

$$\log L \lesssim -\frac{1}{d_{repr}} \log N$$



We show that **dataset intrinsic dimension bounds learned representation intrinsic dimension**:

$$d_{repr} \lesssim d_{data}$$



Want to easily measure these properties of your own datasets?

Check out our code at github.com/mazurowski-lab/intrinsic-properties:

Contact me: @nick_konz
nicholas.konz@duke.edu