REVISITING THE RELATION BETWEEN ROBUSTNESS AND UNIVERSALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

The *modified universality hypothesis* proposed by Jones et al. (2022) suggests that adversarially robust models trained for a given task are highly similar. We revisit the hypothesis and test its generality. We find that predictive behavior does not converge with increasing robustness and thus is not universal. Further, with additional similarity measures, we uncover differences in the representations that were invisible with the measures used in prior work. While robust models tend to be more similar than standard models, robust models remain distinct in important aspects. Moreover, the importance of similarity measures when comparing representations is highlighted as the absolute level of similarity—and thus the assessment of universality—is heavily dependent on the measure used.

1 INTRODUCTION

The universality hypothesis (Olah et al., 2020) suggests that all trained neural networks for a given task are highly similar. If this hypothesis held generally, interpretability research would be simplified, as insights for a specific model could be more easily transferred to other models. While the hypothesis is unlikely to hold in a strict sense (Li et al., 2015; Breiman, 2001), Jones et al. (2022) proposed and presented evidence fo a modified universality hypothesis (MUH): adversarial robustness may function as a strong prior on neural networks such that adversarially robust models will learn similar representations "regardless of exact training conditions (i.e., architecture, random initialization, learning parameters)". They showed empirically that robust CNNs trained on ImageNet (Deng et al., 2009) are highly similar in the input features of the data that they use and in the representations they produce, whereas standard models are not. Thus, training a single robust model is sufficient to mimic the behavior of any other or in their words "if you've trained one, you've trained them all".

However, their work has three key limitations which motivate us to revisit the link between robustness and universality. First, the experiments were centered around representational similarity, while one of the direct and arguably practically most relevant ways to study model similarity is to compare their predictions. Second, a key part of the evidence was gathered with Centered Kernel Alignment (CKA) (Kornblith et al., 2019), a method to measure similarity of representations, which adopts a specific perspective on neural network similarity and was recently shown to have multiple pitfalls (Cui et al., 2022; Dujmović et al., 2022; Davari et al., 2022; Nguyen et al., 2022). Numerous other similarity measures have been proposed (Klabunde et al., 2023; Sucholutsky et al., 2023), which provide alternative views on neural network similarity. Third, experiments exclusively used ImageNet as input data, which leaves the role of data uncertain, e.g., whether results transfer to other vision datasets or out-of-distribution data.

In this work, we thus critically reassess the modified universality hypothesis that suggests that all adversarially robust models for a given task are highly similar. We conduct an extensive empirical study that involves multiple similarity measures, model architectures and datasets. In contrast to previously published results, our study indicates that robust models cannot be considered universal. Our main contributions are:

1. We show that predictions of robust models are not universal (see Figure 1). Their agreement scores do not converge with increasing robustness and the variance of Jensen-Shannon Divergence (JSD) scores increases with higher robustness levels (Section 3.1).

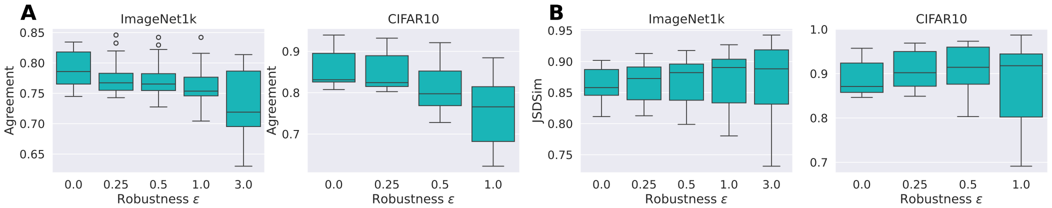

Figure 1: **Predictive behavior remains distinct at high robustness contrary to the MUH.** The boxplots show distributions of agreement (**A**) and scaled Jensen-Shannon divergence (JSDSim) (**B**) across all model pairs when given regular images. The MUH predicts that robust models converge to a universal solution, which should be reflected in highly similar predictions with increasing robustness ϵ. However, predictions do not converge with increasing robustness, with agreement dropping and JSD showing increasing variance. This points towards an issue with the MUH.

2. We observe that representations of robust models are highly similar according to CKA, but not according to other measures like Procrustes or Jaccard similarity. These discrepancies between similarity measures indicate that, on the one hand, adversarial training causes some degree of convergence in the representation space. On the other hand, representations still exhibit critical differences that persist through the final layer and affect model predictions (Section 3.2).
3. We evaluate similarity across different datasets as well as in- and out-of-distribution data. We find that the data has little effect on model similarity. The previously observed phenomena are robust (Section 3.3).

Code and data of our experiments are publicly available (see Appendix D).

2 Background and Methods

Adversarial Robustness While neural networks achieve high performance in many tasks, they are susceptible to —often imperceptible— modifications of inputs that lead to wrong predictions (Szegedy et al., 2014). These modifications δ are usually computed via a constrained optimization problem:

$$\delta^* = \arg\max_{\delta} \mathcal{L}(f(x+\delta), y) \quad \text{s.t.} \quad \|\delta\|_p \leq \epsilon, \tag{1}$$

where $\mathcal{L}$ is the loss function, x, y the input and target, respectively, and ϵ is the strength of the adversarial attack, i.e., the maximal allowed modification of the input. By augmenting training data with adversarial examples, the space of potentially good models is constrained and robust models are produced, which are less susceptible to such attacks (Madry et al., 2019). For these models, perturbations need to be larger to induce misclassifications.

Comparing Predictive Behavior A simple test for universality is comparing predictions of models. If models are universal, we should expect highly similar predictions. Hence, we compare the predicted probability distributions and classifications using average Jensen-Shannon Divergence (JSD) and the agreement rate, respectively. For JSD, we normalize the outputs of the last network layer with a softmax, then compute:

$$\mathrm{JSD}(\boldsymbol{L}, \boldsymbol{L'}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \mathrm{KL}(\boldsymbol{L}_i \| \boldsymbol{L'}_i) + \frac{1}{2} \mathrm{KL}(\boldsymbol{L'}_i \| \boldsymbol{L}_i), \tag{2}$$

where $\boldsymbol{L}, \boldsymbol{L'} \in \mathbb{R}^{N \times C}$ are the collections of the predicted class probabilities, i.e., the softmaxed logits, for C classes and N fixed inputs, and KL is the Kullback-Leibler Divergence. In the rest of the paper, we report JSDSim, i.e., scaled and normalized JSD to the range of $[0, 1]$, such that a score of 1 indicates identical predicted distributions.

The agreement rate is the rate of instances that are predicted as the same class. This can be notated as the argmax of the logits, with $\mathbf{1}[\cdot]$ as the indicator function:

$$\text{Agreement}(\boldsymbol{L}, \boldsymbol{L'}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\arg\max_j \boldsymbol{L}_{ij} = \arg\max_j \boldsymbol{L'}_{ij}]. \tag{3}$$

Comparing Representations Another approach at testing universality is comparing the internal representations, i.e., the activation of a layer for some input. Again, if models are universal, we expect that their internal processes are highly similar, which should lead to similar representations. To measure representational similarity, activations are collected for a set of inputs resulting in a matrix $\boldsymbol{R} \in \mathbb{R}^{N \times D}$, where N is the number of inputs and D is the number of neurons in the layer. A representational similarity measure typically takes two such matrices as input and produces a single number that quantifies the similarity of these matrices. The matrices may come from different layers or models, but are based on the same set of inputs such that the rows between the matrices correspond. The similarity score respects certain transformations between representations that would keep them equivalent, e.g., switching neuron order, which would result in a different order of the columns of the compared matrices. For a detailed introduction, we refer to the survey by Klabunde et al. (2023).

In this work, we use three similarity measures: linear CKA (Kornblith et al., 2019), Orthogonal Procrustes (Procrustes) (Ding et al., 2021; Williams et al., 2021), and k-NN Jaccard Similarity (Jaccard). Intuitively, these measures summarize the similarity of representations across multiple different aspects, e.g., specific properties of their geometry. These measures have been empirically shown to give meaningful similarity assessments (Klabunde et al., 2024), but highlight different discrepancies between representations. Thus, employing a set of similarity measures enables a more multi-faceted comparison of representations. At the same time, the measures we use consider the same representations equivalent, i.e., any representations that only differ in rotation, reflection, scale, and translation. This means we should expect similar similarity scores when representations are close to equivalent.

Formally, CKA computes a similarity score between 0 and 1 given two centered representations $\boldsymbol{R} \in \mathbb{R}^{N \times D}, \boldsymbol{R'} \in \mathbb{R}^{N \times D'}$, i.e., with zero mean columns, as follows:

$$\text{CKA}(\boldsymbol{R}, \boldsymbol{R'}) = \frac{\|\boldsymbol{R'}^{\mathsf{T}} \boldsymbol{R}\|_F^2}{\|\boldsymbol{R}^{\mathsf{T}} \boldsymbol{R}\|_F \|\boldsymbol{R'}^{\mathsf{T}} \boldsymbol{R'}\|_F}, \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm. Based on the overall feature correlations, CKA measures global representational similarity.

Procrustes is another measure with a global view on similarity and satisfies the criteria of a metric. Procrustes finds the optimal orthogonal alignment between two representation spaces:

$$\text{Procrustes}(\boldsymbol{R}, \boldsymbol{R'}) = \min_{\boldsymbol{Q}} \|\boldsymbol{R}\boldsymbol{Q} - \boldsymbol{R'}\|_F = (\|\boldsymbol{R}\|_F^2 + \|\boldsymbol{R'}\|_F^2 - 2\|\boldsymbol{R}^{\mathsf{T}}\boldsymbol{R'}\|_*)^{1/2}, \tag{5}$$

where $\|\cdot\|_*$ is the nuclear norm, i.e., the sum of the singular values. As $\boldsymbol{R}, \boldsymbol{R'}$ need to have equal dimension for Procrustes, we zero-pad the representation with lower dimension. In addition to zero-centering the columns, we scale the representation matrix to unit norm. With this, we report $\frac{2-\text{Procrustes}}{2}$ as ProcrustesSim, which is scaled to $[0, 1]$, where 1 indicates maximal similarity. Finally, we use Jaccard for a local view on representation similarity. Jaccard is defined as the average intersection over union of the nearest neighbors in the representation spaces:

$$\text{Jaccard}(\boldsymbol{R}, \boldsymbol{R'}) = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathcal{N}_i^k(\boldsymbol{R}) \cap \mathcal{N}_i^k(\boldsymbol{R'})|}{|\mathcal{N}_i^k(\boldsymbol{R}) \cup \mathcal{N}_i^k(\boldsymbol{R'})|}, \tag{6}$$

where $\mathcal{N}_i^k(\boldsymbol{R})$ are the k nearest neighbors of the representation of input i in $\boldsymbol{R}$. We use $k = 10$ and cosine similarity on the centered representations to find the nearest neighbors.

Detecting Differences in the Representation Mechanism with Image Inversion One problem of similarity measures is that they do not pick up on differences in the usage of input features as long as models produce similar representations or predictions (Jones et al., 2022). This may lead to overestimation of similarity between two neural networks. We thus aim to test the similarity of

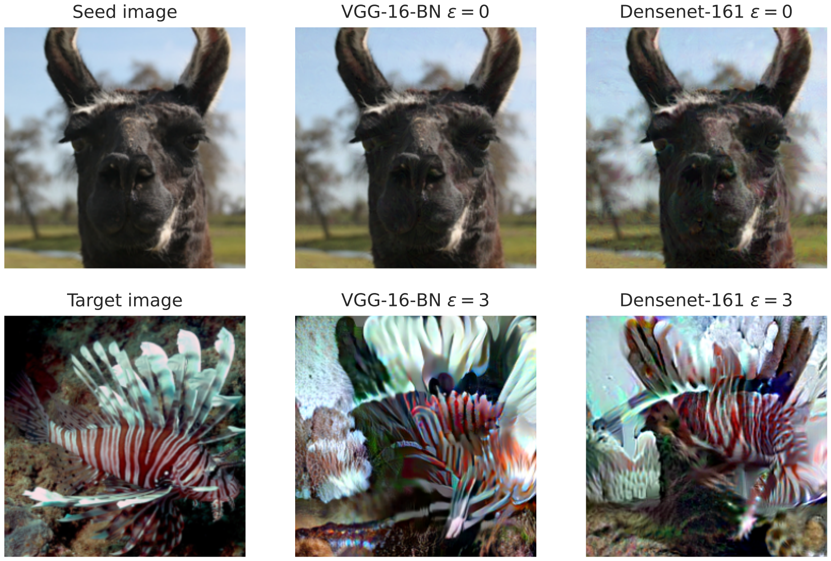
Seed image
VGG-16-BN $\varepsilon = 0$
Densenet-161 $\varepsilon = 0$
Target image
VGG-16-BN $\varepsilon = 3$
Densenet-161 $\varepsilon = 3$

Figure 2: **Example of seed, target, and inverted image.** Shown for VGG16 (left) and DenseNet161 (right) trained in ImageNet1k. The top rows shows results for standard models ($\epsilon = 0$), the bottom for robust models ($\epsilon = 3$). The inverted images produced by robust and standard models are quite different. Inverted images of standard models are visually extremely similar to the seed image. For robust models, inverted images contain elements clearly belonging to the target image. They show how feature cooccurrence can be lessened, e.g., the background of the fish was not added to the image as both models mainly rely on the fins and texture of the fish.

the combination of the input feature reliance and the processing into a representation. We call this combination the *representation mechanism.*

Image inversion (Ilyas et al., 2019) presents a way to create a model-specific variant of an input that produces nearly the same representation as the original input, but only consistently contains the input features that the model actually uses. Hence, *inverted images* enable the study of the similarity of representation mechanisms. If one model has a different mechanism that relies on another set of input features, it will not find those features in the inverted image of another model and thus will be unable to produce a similar representation. Comparing the representations given inverted images gives us information about the similarity of the mechanisms.

To create an inverted image $\tilde{\boldsymbol{x}}$ for a given target image $\boldsymbol{x}$, a seed image $\boldsymbol{s}$ from a different class is modified such that it produces a representation similar to the representation of the target image. More precisely, let $f^L(\boldsymbol{x}) \in \mathbb{R}^D$ be the representation of model f for the target image $\boldsymbol{x}$ in the penultimate layer L, then the inverted image $\tilde{\boldsymbol{x}}$ is computed as the output of

$$\min_{\boldsymbol{s}} \frac{\|f^L(\boldsymbol{s}) - f^L(\boldsymbol{x})\|_2}{\|f^L(\boldsymbol{x})\|_2}. \tag{7}$$

The optimization is done with gradient descent, so the naive solution of $s = x$ is not reached. Instead, the most relevant input features for the model f are introduced to the seed image. As the seed image is sampled randomly from all images with a different class than the target image, feature cooccurrence in natural images, e.g., dog fur texture and dog ears, can be eliminated if only of those features is relevant for f. See Figure 2 for an example.

3 Experiments

We will now lay out the general setup for the experiments to test the MUH.

Models We initially use L_2-robust models trained on ImageNet-1k (Deng et al., 2009) and CIFAR-10 (Krizhevsky, 2009). The full list of models is given in Appendix A. While we train most of these

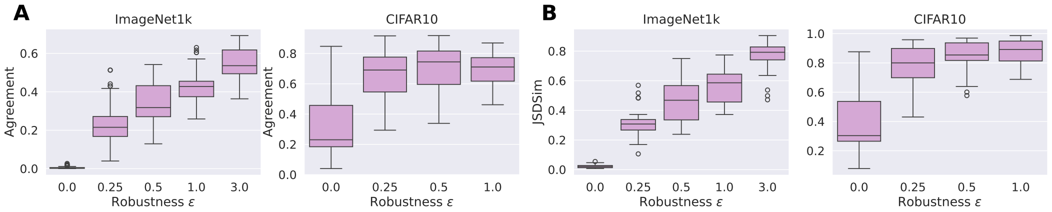

Figure 3: **Similarity of predictions on inverted images increases with robustness.** The boxplots show the agreement (**A**) and JSDSim (**B**) distribution across all model pairs when given inverted images. Both agreement and JSDSim increase with increasing robustness on both ImageNet1k and CIFAR-10. This means that robustness does lead to increased similarity in some aspects of the models, but arguably not to universality as the absolute similarity remains relatively low.

models ourselves, we use the checkpoints released by Salman et al. (2020) for ImageNet-1k. We study models with robustness of $\epsilon \in \{0, 0.25, 0.5, 1.0, 3.0\}$ on ImageNet-1k, but stop at $\epsilon = 1$ for CIFAR-10 due to the lower resolution of images.

General Setup We compare models either using regular images or inverted images as input. For convenience, figures have color schemes corresponding to the type of input. As inverted images are model specific, each pair of models A, B is compared given inverted images of A and of B. All comparisons are made within one level of robustness, i.e., A and B always were trained with the same ϵ. Comparisons are made with multiple similarity measures, as outlined in Section 2. For representational similarity, we collect the model activations at the second to last, or penultimate, layer.

3.1 Predictions of Robust Models Not Universal

If adversarially robust models are universal in a strict sense, we would expect that their predictions overlap to a very high degree. Figure 1A shows that this is not the case. On regular images, the agreement between predictions of highly robust models is much lower than the theoretical maximum imposed by small accuracy differences (Fort et al., 2019) and comparable to the agreement between standard models. Comparing the predicted distributions with Jensen-Shannon divergence (Figure 1B) instead of just the final predictions leads to the same conclusion: the predictive behavior is not universal.

However, using inverted images as input, which highlight differences in the representation mechanism, reveals that robustness does have a profound impact on similarity of models. Figure 3 shows how predictive behavior on these kind of inputs become more similar with increasing robustness. Jones et al. (2022) showed similar effects for similarity of the representations directly. Nevertheless, the differences in predictions given regular data must have an origin. In the following section, we will present a possible explanation.

3.2 Differences in Representation Mechanisms of Robust Models

Closely connected to the final predictions are the representations at the penultimate layer of the neural network as they are the input for the final classification layers. Should these representations be highly similar, we intuitively expect that the classification layers will lead to similar predictions. Conversely, if the representations are not similar, similar predictions might still be possible, but we expect them to be less likely. Hence, inspecting the representations may lead to an explanation why predictions of models are different.

Using inverted images as inputs lets us study similarity of the whole representation mechanism, i.e., the combination of input feature reliance and feature processing, as mentioned in Section 2. If this mechanism is similar, predictions should be similar as well as the only degrees of freedom remain in the final classifier layers. Jones et al. (2022) found high similarity between the representations of inverted images using CKA, a highly popular similarity measure. We now compare the representation

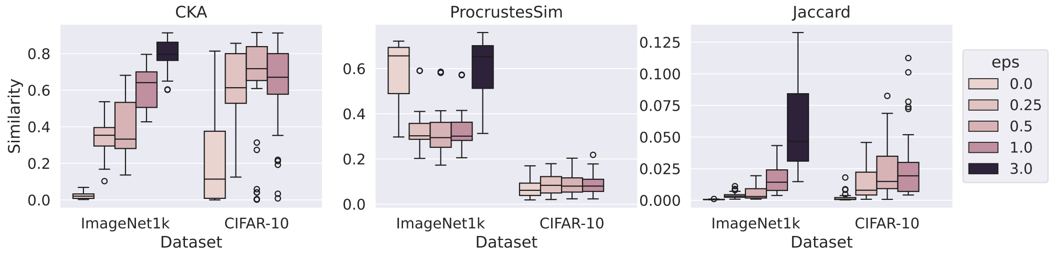

Figure 4: **Multiple perspectives on representation mechanism similarity reveal differences.** Distributions of representational similarity within each robustness level given inverted images as input. Each group of boxes corresponds to one dataset, with robustness increasing from left to right. Each panel shows the results for a different similarity measure. While CKA increases steadily with increased robustness, the effect on ProcrustesSim is minor. Jaccard similarity increases slowly but remains low on an absolute level even with high robustness. Its values indicate that there is no overlap between the ten nearest neighbors on average as it is less than 0.05 in most cases. ProcrustesSim reveals the existence of differences in global structure of the representations that CKA does not recognize. Jaccard additionally highlights differences in local structure, which cannot be detected by a global measure like linear CKA.

mechanism using two additional measures, namely ProcrustesSim and Jaccard similarity, to get a more comprehensive view on the similarity of representation mechanisms.

Figure 4 shows that robustness does not lead to highly similar representation mechanisms from all perspectives. Instead, the additional ProcrustesSim and Jaccard similarity measures suggest that the absolute level of similarity is relatively low and increases only little with higher levels of robustness. In particular, ProcrustesSim barely changes on CIFAR-10 across all robustness levels. On ImageNet1k, ProcrustesSim is higher, but still far from 1 and displays a U-like shape. We do not have an explanation for the shape—the models are taken from Salman et al. (2020) and were trained identically across ϵ levels[1]. In contrast, the CIFAR-10 models, which we trained ourselves, do not display such a pattern. As the pattern does not have a major impact on our conclusions regarding universality, we regard investigating the origin of the pattern as an interesting question for future work. Jaccard similarity shows that on average there is no overlap between the ten nearest neighbors in representation space. Appendix B presents results with other neighborhood sizes; overlap remains small.

Overall, our results suggest that adversarial training causes robust representations to converge to some extent. However, the representations still maintain differences which some similarity measures fail to detect. This demonstrates a pitfall when analyzing model similarity by relying only on a single measure as well as the importance of a multifaceted analysis.

3.3 How Robust is the Effect of Robustness on Similarity?

So far, we have shown on ImageNet1k and CIFAR-10 that robustness increases similarity of neural networks in some aspects, but also that increasing robustness does not eliminate all differences between models. However, the degree of model similarity might be impacted by the training data and we have yet to examine the influence of data in detail. On the one hand, larger datasets might lead to increased model similarity as more data could lead to less overfitting. On the other hand, larger datasets tend to be more complex, i.e., contain more classes as well as a larger variety of images, which could lead to a more complex loss landscape that makes it difficult to converge to similar optima. Additionally, out-of-distribution (OOD) behavior might differ significantly from in-distribution behavior and is particularly well suited to highlight differences between models (Ding et al., 2021).

We thus evaluate the influence of data on model similarity in two experiments. First, we compare models on three variations of ImageNet, i.e., ImageNet1k, ImageNet100 and ImageNet50, to assess

[1] However, we found that checkpoint metadata like the epoch number was inconsistent with the performance of some models hinting at a potential issue.

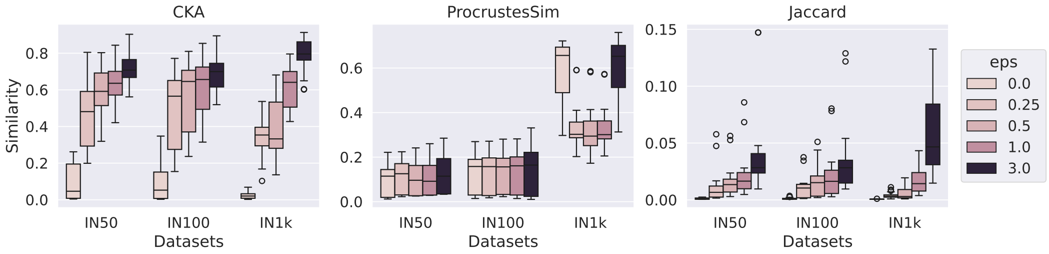

Figure 5: **Effect of dataset size on representation mechanism similarity across robustness levels.** We test the influence of the training data on model similarity by repeating the experiment of Section 3.2 across models trained on ImageNet subsets with 50 and 100 randomly sampled classes. For the comparisons, models use inverted images. Each group of boxes corresponds to the results for one dataset, with robustness increasing from left to right. Observations are generally similar across datasets, i.e., CKA and Jaccard increase, whereas ProcrustesSim remains roughly constant over robustness levels. Differences are that CKA rises more slowly on ImageNet1k and that it does not seem to saturate. Further, the absolute level on ProcrustesSim increases slightly with dataset size. The U-shape on Imagenet1k is absent on the other datasets, but we hypothesize that the training procedure of those models was not identical. Hence, the diversity of the data seems to have only a small effect on similarity of robust models.

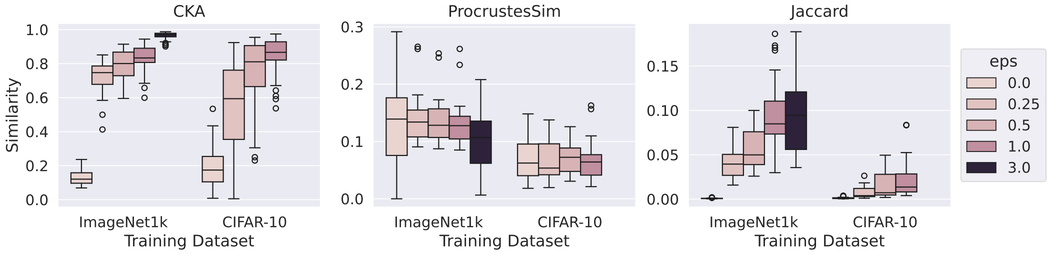

Figure 6: **Effect of OOD data on representation mechanism similarity across robustness levels.** We use inverted images generated with SAT-6 as inputs for ImageNet1k and CIFAR-10 models to test the effect of OOD data on representational similarity. Each group of boxes shows the results for one dataset, with robustness increasing from left to right. Increasing robustness has similar effects across datasets, even though the scales of similarity scores differ. Overall, similarity on OOD data is higher than on in-distribution images.

the effect of dataset size and complexity. ImageNet50 and ImageNet100 contain a sample of the full ImageNet1k classes, but are otherwise identical. Second, we evaluate how out-of-distribution data impacts model similarity using the SAT-6 dataset (Basu et al., 2015) as input. SAT-6 consists of 28x28 pixel airborne images of California, thus image content and perspective are OOD for models trained on ImageNet1k or CIFAR-10.

Effect of Dataset Size Figure 5 shows the similarity of the representation mechanisms given inverted images as inputs. Across the ImageNet variations, the trends are similar. CKA increases quickly with increasing robustness, ProcrustesSim stays roughly constant, and Jaccard Similarity increases slowly. Notably, on the smaller datasets, CKA increases more quickly compared to ImageNet1k. However, the range of similarity scores remains large initially. Additionally, ProcrustesSim is lower on the small datasets. Hence, the amount of data seems to have some but little influence on the similarity of robust models.

Effect of OOD Data Figure 6 shows representation mechanism similarity on inverted images that were generated using the SAT-6 dataset. Both ImageNet1k and CIFAR-10 models exhibit similar trends on OOD data. CKA and Jaccard Similarity increase with robustness and are higher than the scores on inverted images from the respective dataset reported in Figure 4. In contrast, ProcrustesSim scores are lower for both CIFAR-10 and ImageNet1k.

All in all, these experiments show that while the MUH does not hold completely, robustness leads to consistently increased similarity of models in some aspects. The phenomenon is robust across dataset size and OOD data.

4 Discussion

MUH Only Holds Partially With our experiments, we demonstrated that predictions of robust models do not converge with increasing robustness and highlighted the existence of differences in the representation mechanism. Hence, the MUH does not hold in full. However, consistent with Jones et al. (2022), we also observed that representation mechanisms consistently become more similar with increased robustness, when measured by CKA and Jaccard—albeit slowly. While more work is needed to understand the origin of the observations in detail, they point towards an interesting direction of future work: why is it that some aspects of models seem to be strongly constrained by robustness whereas others are not?

(Un-)Interpretability of Representational Similarity Measures As we found in Section 3.2, Procrustes highlights that there is some difference in the representations from inverted images. However, in contrast to interpretable measures like agreement on predictions or Jaccard on representations, the score of Procrustes is hard to interpret. While we know from the definition of the similarity measure that representations must be unable to be aligned well, what exactly makes this alignment difficult is unclear. The similarity scores point us towards model pairs that appear dissimilar and is thus useful as a first step, but ultimately gives limited insights into the representation space. We conducted some exploratory analysis, but finding the origin of a low representational similarity score is not straight forward. At the same time, the observation is unlikely to be a defect of the similarity measure as it performs well in benchmarking tests (Klabunde et al., 2024). This may highlight a fundamental problem of representational similarity measures trying to break down similarity between objects with many facets into a single number.

Impact on Interpretability Research On the one hand, the observed differences point out a attack point for future work to understand both robust models and representational similarity measures better. This could lead to more effective use of similarity measures as explorative tools. On the other hand, our work is another point of evidence against universality in a strong sense, where all parts of a model are highly similar, and towards a world where models consist of universal and non-universal parts. Studying universal parts may be of general interest, whereas non-universal parts may be only interesting for frontier models or specific models with high interest. Hence, identifying universal components is an interesting direction of future work.

Increasing Robustness Beyond Our Experiments We observed that increasing robustness up to $\epsilon = 3$ for ImageNet models lead to increased similarity in some aspects of the models, e.g., the trend for CKA similarity appears to continue further. Thus, extremely robust models may be a way to studying the whole model class at once–at least with respect to the aspects that make them similar from the CKA perspective. However, increasing robustness even further would likely lead to further accuracy degradation. Ultimately, such models may be not comparable to more widely used models, which could make detailed study of these models not worth it despite the aforementioned benefit.

Value of Multi-Faceted Similarity Analysis Our work demonstrates that reliance on a single measure when evaluating model similarity can lead to over- or underestimation of similarity. Combining similarity measures and analyzing predictions as well as representations provides a more comprehensive view of relations between models. Thus, the similarity assessment is less prone to rely on shortcomings of individual measures. While this approach comes at a (small) cost—the popular CKA is very fast compared to other measures—we argue that the trade-off is worth it.

Exchanging Models in Production When a model that delivers predictions to users should be exchanged, e.g., by a more efficient variant, consistency of predictions may be desired (Milani Fard et al., 2016a). The MUH suggested an easy approach to increase prediction consistency by using (slightly) adversarially robust models. As we find that predictions do not become more similar, other methods to improve consistency are necessary.

5 Related Work

Universality The question to what extent models are universal has attracted significant interest in prior work. On the one hand, model multiplicity, i.e., the existence of multiple models with almost equal performance but different input-output behavior or representations, has been studied extensively (Breiman, 2001; Black et al., 2022; Heljakka et al., 2023). Architecturally similar models trained or updated on near-identical data can differ significantly (Klabunde and Lemmerich, 2023; Somepalli et al., 2022; Marx et al., 2020; Black and Fredrikson, 2021; Liu et al., 2022; McCoy et al., 2020; Li et al., 2015). Modifications to training or inference may be necessary to enforce consistent behavior between different models (Milani Fard et al., 2016b; Summers and Dinneen, 2021). In mechanistic interpretability, a more fine-grained view on universality is taken, i.e., whether the input-output behavior of a network is also implemented in the same way. It leads to further evidence against universality (Zhong et al., 2023; Chughtai et al., 2023).

On the other hand, there is evidence for universality in certain scenarios. Some features consistently appear in CNNs (Schubert et al., 2021). Further, attention heads with specific functionality can be found across many transformer-based language models (Olsson et al., 2022; Gould et al., 2023). Additionally, some of their internal processes for tasks such as indirect object identification (Merullo et al., 2023) and retrieval (Variengien and Winsor, 2023) seem to be universal, at least across certain model classes. On the smallest scale, certain neurons appear universal (Gurnee et al., 2024). Further, parts of two different models (trained for the same task) can be connected using simple transformations with little accuracy loss (Csiszárik et al., 2021; Bansal et al., 2021; Lähner and Moeller, 2023; Moschella et al., 2023) indicating representational compatibility (Brown et al., 2023).

While the two above collections of evidence for and against universality might seem contradicting, the scope of universality as well as what would be considered equivalent between networks differs drastically. In fact, universality has multiple non-binary facets (Gurnee et al., 2024). Further, universality may only occur for certain types of models (Jones et al., 2022).

Neural Network Similarity To measure similarity of neural networks, especially of their representations, numerous similarity measures have been proposed across machine learning and neuroscience (Klabunde et al., 2023; Sucholutsky et al., 2023). These measures represent different views on what kind of behavior is considered equivalent. Due to its popularity, Centered Kernel Alignment (CKA) (Kornblith et al., 2019) has attracted particular interest and was also used by Jones et al. (2022) who propose the hypothesis of universality across robust models. However, several caveats of CKA are known: few data points may dominate the similarity score (Nguyen et al., 2022), the choice of inputs may determine similarity measurements in early layers (Cui et al., 2022), and scores are generally brittle (Davari et al., 2022).

6 Conclusion

We revisit the modified universality hypothesis which states that adversarially trained models are highly similar. We show that predictions of robust models are not universal as their agreement on regular images does not converge with robustness. Additionally, the similarity of representation mechanisms, i.e., the combination of input feature reliance and processing of the input into a representation, is highly dependent on the way similarity is measured. While CKA shows that robust representation mechanisms are highly similar, ProcrustesSim does not increase with robustness. Our multi-faceted similarity analysis reveals that while adversarial training causes robust models to converge to some degree, they still exhibit critical differences that some similarity measures fail to detect, but affect predictive behavior. Our results show that the modified universality hypothesis in its original form does not hold and highlights the importance of using a broader set of measures when evaluating model similarity.

Reproducibility Statement

All code and data to reproduce our results are publicly available, see Appendix D for details.

REFERENCES

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024.001, March 2020. ISSN 2476-0757. doi:10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent Learning: Do different neural networks learn the same representations? In Dmitry Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar, editors, *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings of Machine Learning Research*, pages 196–212, Montreal, Canada, 11 Dec 2015. PMLR. URL https://proceedings.mlr.press/v44/li15convergent.html.

Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, August 2001. ISSN 0883-4237, 2168-8745. doi:10.1214/ss/1009213726. URL https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full. Publisher: Institute of Mathematical Statistics.

Haydn T. Jones, Jacob M. Springer, Garrett T. Kenyon, and Juston S. Moore. If you've trained one you've trained them all: inter-architecture similarity increases with robustness. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 928–937. PMLR, 01–05 August 2022. URL https://proceedings.mlr.press/v180/jones22a.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi:10.1109/CVPR.2009.5206848.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR, May 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html. ISSN: 2640-3498.

Tianyu Cui, Yogesh Kumar, Pekka Marttinen, and Samuel Kaski. Deconfounded Representation Similarity for Comparison of Neural Networks. *Advances in Neural Information Processing Systems*, 35:19138–19151, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/79cbf4f96c2bcc67267421154da689dd-Abstract-Conference.html.

Marin Dujmović, Jeffrey S. Bowers, Federico Adolfi, and Gaurav Malhotra. The pitfalls of measuring representational similarity using representational similarity analysis, April 2022. URL https://www.biorxiv.org/content/10.1101/2022.04.05.487135v1.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. On the Inadequacy of CKA as a Measure of Similarity in Deep Learning. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022. URL https://openreview.net/forum?id=rK841rby6xc.

Thao Nguyen, M. Raghu, and Simon Kornblith. On the Origins of the Block Structure Phenomenon in Neural Network Representations. *Trans. Mach. Learn. Res.*, February 2022. URL https://www.semanticscholar.org/paper/On-the-Origins-of-the-Block-Structure-Phenomenon-in-Nguyen-Raghu/5fe4f6fbe26f94ff65290c58007185ec71669921.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of Neural Network Models: A Survey of Functional and Representational Measures, August 2023. URL http://arxiv.org/abs/2305.06329.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, November 2023. URL http://arxiv.org/abs/2310.13018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, February 2014. URL http://arxiv.org/abs/1312.6199.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*. arXiv, September 2019. URL http://arxiv.org/abs/1706.06083.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding Representation Similarity Through Statistical Testing. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1556–1568. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0c0bf917c7942b5a08df71f9da626f97-Paper.pdf.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized Shape Metrics on Neural Representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4738–4750. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/252a3dbaeb32e7690242ad3b556e626b-Paper.pdf.

Max Klabunde, Tassilo Wald, Tobias Schumacher, Klaus Maier-Hein, Markus Strohmaier, and Florian Lemmerich. Resi: A comprehensive benchmark for representational similarity measures. *arXiv preprint arXiv:2408.00531*, 2024.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e2c420d928d4bf8ce0ff2ec19b371514-Paper.pdf.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. page 60, 2009.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do Adversarially Robust ImageNet Models Transfer Better? In *Advances in Neural Information Processing Systems*, volume 33, pages 3533–3545. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/24357dd085d2c4b1a88a7e0692e60294-Abstract.html.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10, 2015. URL https://dl.acm.org/doi/abs/10.1145/2820783.2820816.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and iterate: Reducing prediction churn. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016a. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/dc5c768b5dc76a084531934b34601977-Paper.pdf.

Emily Black, Manish Raghavan, and Solon Barocas. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi:10.1145/3531146.3533149. URL https://dl.acm.org/doi/10.1145/3531146.3533149.

Ari Heljakka, Martin Trapp, Juho Kannala, and Arno Solin. Disentangling Model Multiplicity in Deep Learning, January 2023. URL http://arxiv.org/abs/2206.08890.

Max Klabunde and Florian Lemmerich. On the Prediction Instability of Graph Neural Networks. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 187–202. Springer Nature Switzerland, 2023. ISBN 978-3-031-26409-2. doi:10.1007/978-3-031-26409-2_12.

Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. Can Neural Nets Learn the Same Model Twice? Investigating Reproducibility and Double Descent from the Decision Boundary Perspective. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13689–13698, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi:10.1109/CVPR52688.2022.01333. URL https://ieeexplore.ieee.org/document/9878514/.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive Multiplicity in Classification. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6765–6774. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/marx20a.html.

Emily Black and Matt Fredrikson. Leave-one-out Unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 285–295, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi:10.1145/3442188.3445894. URL https://dl.acm.org/doi/10.1145/3442188.3445894.

Huiting Liu, Avinesh P. V. S., Siddharth Patwardhan, Peter Grasch, and Sachin Agarwal. Model Stability with Continuous Data Updates. *arXiv:2201.05692 [cs]*, January 2022. URL http://arxiv.org/abs/2201.05692.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.blackboxnlp-1.21. URL https://aclanthology.org/2020.blackboxnlp-1.21.

Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and Iterate: Reducing Prediction Churn. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016b. URL https://proceedings.neurips.cc/paper/2016/hash/dc5c768b5dc76a084531934b34601977-Abstract.html.

Cecilia Summers and Michael J. Dinneen. Nondeterminism and Instability in Neural Network Optimization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9913–9922. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/summers21a.html.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 27223–27250. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/56cbfbf49937a0873d451343ddc8c57d-Paper-Conference.pdf.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A Toy Model of Universality: Reverse Engineering How Networks Learn Group Operations. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6243–6267. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/chughtai23a.html`.

Ludwig Schubert, Chelsea Voss, Nick Cammarata, Gabriel Goh, and Chris Olah. High-Low Frequency Detectors. *Distill*, 2021. doi:10.23915/distill.00024.005.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context Learning and Induction Heads. 2022. URL `https://arxiv.org/abs/2209.11895`.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor Heads: Recurring, Interpretable Attention Heads In The Wild. In *The Twelfth International Conference on Learning Representations*, October 2023. URL `https://openreview.net/forum?id=kvcbV8KQsi`.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit Component Reuse Across Tasks in Transformer Language Models. In *The Twelfth International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=fpoAYV6Wsk`.

Alexandre Variengien and Eric Winsor. Look Before You Leap: A Universal Emergent Decomposition of Retrieval Tasks in Language Models. 2023. doi:10.48550/ARXIV.2312.10091. URL `https://arxiv.org/abs/2312.10091`.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal Neurons in GPT2 Language Models, January 2024. URL `http://arxiv.org/abs/2401.12181`.

Adrián Csiszárik, Péter Kőrösi-Szabó, Ákos Matszangosz, Gergely Papp, and Dániel Varga. Similarity and Matching of Neural Network Representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5656–5668. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/2cb274e6ce940f47beb8011d8ecb1462-Paper.pdf`.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting Model Stitching to Compare Neural Representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 225–236. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/01ded4259d101feb739b06c399e9cd9c-Paper.pdf`.

Zorah Lähner and Michael Moeller. On the Direct Alignment of Latent Spaces. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, December 2023. URL `https://openreview.net/forum?id=nro8tEfIfw`.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, February 2023. URL `https://openreview.net/forum?id=SrC-nwieGJ`.

Davis Brown, Charles Godfrey, Nicholas Konz, Jonathan Tu, and Henry Kvinge. Understanding the Inner Workings of Language Models Through Representation Dissimilarity, October 2023. URL `http://arxiv.org/abs/2310.14993`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2017.

Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. TinyViT: Fast Pretraining Distillation for Small Vision Transformers. In *European Conference on Computer Vision*, pages 68–85, 2022. URL https://dl.acm.org/doi/abs/10.1007/978-3-031-19803-8_5.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

Table 1: The number of parameters and accuracy (Acc) and adversarial accuracy (Adv. Acc.) for models trained on ImageNet1k. The adversarial accuracy of models with $\epsilon = 0$ was evaluated with $\epsilon = 0.25$. For the models marked in gray, we used the checkpoints provided by Salman et al. (2020).

Architectures	Parameters	$\epsilon = 0$		$\epsilon = 0.25$		$\epsilon = 0.5$		$\epsilon = 1$		$\epsilon = 3$	
		Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.
ResNet-18	11.7M	69.80	20.30	67.42	60.02	65.48	55.97	62.31	55.65	53.11	49.70
ResNet-50	25.6M	75.80	25.97	74.13	67.42	73.17	64.23	70.42	64.32	62.83	59.47
Wide ResNet-50-2	68.9M	76.98	29.37	76.22	69.82	75.11	66.70	73.42	67.36	66.90	63.45
Wide ResNet-50-4	223.4M	77.91	32.74	77.10	72.82	76.52	69.00	75.51	62.78	69.67	45.17
ResNeXt-50 32x4d	28.7M	77.32	26.00	-	-	59.74	49.73	72.45	66.71	65.92	62.39
Densenet-161	25.0M	77.38	28.78	-	-	-	-	60.12	13.33	66.12	62.72
VGG-16-BN	138.4M	73.67	10.86	68.49	61.57	68.32	59.29	66.33	60.14	56.79	53.51
TinyViT	5M	72.65	24.03	71.16	65.27	69.30	60.87	66.49	60.58	56.45	53.22

A Additional Model Information

A.1 ImageNet1k Models

Table 1 shows all model architectures with their accuracy and number of parameters. We use seven L_2-robust CNNs: ResNet-18, ResNet-50 (He et al., 2016), Wide ResNet-50-2, Wide ResNet-50-4 (Zagoruyko and Komodakis, 2017), ResNeXt-50 32x4d (Xie et al., 2017), Densenet-161 (Huang et al., 2017), and VGG-16-BN (Simonyan and Zisserman, 2015) along with a TinyViT (Wu et al., 2022) trained from scratch.

Training Details (Salman et al., 2020) trained their L_2-robust ImageNet models for 90 epochs using an initial learning rate of 0.1 which is reduced every 30 epochs by a factor of 10. The training uses stochastic gradient descent (SGD) with a batch size of 512, a momentum of 0.9 and weight decay of $1e^{-4}$. For standard training, cross-entropy was used as a loss function. Robust training was conducted using projected gradient descent (PGD) (Madry et al., 2019) allowing L_2 perturbation of the respective ϵ value. Adversarial examples were generate in three attack steps with a step size of $\frac{2}{3}\epsilon$. The TinyViT models were trained with the same setup but with a lower batch size of 256. We used an identical setting for training the remaining ImageNet1k models.

Inverted Images Inverted images were generated on the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) validation set using the Robustness library (Engstrom et al., 2019). First 10,000 target images were randomly sampled from the dataset. Then, for each sampled image, a seed image was sampled at random. If the sampled seed image had the same class as the target, a new image was sampled until seed and target classes were different. To generate an inverted image, the seed image was modified in three steps and the best result taken.

A.2 ImageNet100 Models

Table 2 shows accuracy scores for ImageNet100 models.

Training Details We trained the ImageNet100 models using the same training procedure as for ImageNet.

Inverted Images The process for generating inverted images is identical to that on ImageNet. Seed and target images were sampled from the ImageNet100 train set.

Table 2: The number of parameters and accuracy (Acc) and adversarial accuracy (Adv. Acc.) for models trained on ImageNet100. The adversarial accuracy of models with $\epsilon = 0$ was evaluated with $\epsilon = 0.25$.

Architectures	Parameters	$\epsilon = 0$		$\epsilon = 0.25$		$\epsilon = 0.5$		$\epsilon = 1$		$\epsilon = 3$	
		Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.
ResNet-50	25.6M	79.00	45.92	79.28	73.60	77.16	68.44	74.30	60.86	69.88	47.54
Wide ResNet-50-2	68.9M	80.50	51.44	80.22	74.88	79.92	72.40	75.64	63.98	69.22	46.56
Densenet-161	25.0M	83.30	57.48	83.24	78.32	82.16	75.22	81.20	69.60	76.26	54.24
VGG-16-BN	138.4M	82.52	41.36	80.02	74.56	78.62	69.46	73.86	62.06	66.46	45.76

Table 3: The number of parameters and accuracy (Acc) and adversarial accuracy (Adv. Acc.) for models trained on ImageNet50. The adversarial accuracy of models with $\epsilon = 0$ was evaluated with $\epsilon = 0.25$.

Architectures	Parameters	$\epsilon = 0$		$\epsilon = 0.25$		$\epsilon = 0.5$		$\epsilon = 1$		$\epsilon = 3$	
		Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.
ResNet-50	25.6M	72.32	40.80	78.32	72.80	76.92	69.28	74.76	63.20	67.84	49.16
Wide ResNet-50-2	68.9M	76.76	48.44	77.36	72.04	76.04	68.24	75.52	64.84	70.52	50.64
Densenet-161	25.0M	81.28	58.24	81.56	76.00	80.20	72.40	78.20	68.36	73.36	54.12
VGG-16-BN	138.4M	73.80	40.96	80.80	75.88	76.32	68.84	71.48	59.16	65.60	45.60

A.3 IMAGENET50

Dataset Creation The dataset was created by sampling a random subset of 50 classes from ImageNet100 and taking all images of the chosen classes.

Training Details The training procedure for ImageNet50 models was the same as for ImageNet1k with a few deviations. As models converged faster on the smaller datasets, we reduced the epochs to 60 with the exception of VGG-16-BN for $\epsilon = 1$ and $\epsilon = 0.5$. As the VGG-16-BN models with $\epsilon = 3$ and $\epsilon = 0.25$ initially did not converge, we reduced the learning rate to 0.05 and trained them with batch size 256. Due to availability of hardware, other models were trained with batch size 128 for regular training, except DenseNet-161 with batch size 64. For adversarial training, batch size was reduced to 64 and 32, respectively.

Inverted Images The process for generating inverted images is identical to that on ImageNet. Seed and target images were sampled from the ImageNet50 train set.

A.4 CIFAR-10 MODELS

Table 4 shows the accuracy and number of parameters of each CIFAR-10 CNN.

Training Details The CIFAR-10 models were trained using almost the same configuration as the L_2-robust ImageNet1k CNNs. The only modification for standard training was using a weight decay of $5e^{-4}$.

Inverted Images Seed and target images were taken from the CIFAR-10 test set, which contains 10,000 images.

B JACCARD SIMILARITY WITH VARYING NEIGHBORHOOD SIZE

Figure 7 shows additional result for Jaccard similarity with neighborhood sizes $k \in \{10, 100, 500\}$.

Table 4: The number of parameters, accuracy (Acc.) and adversarial accuracy (Adv. Acc.) for models trained on CIFAR-10. The adversarial accuracy of models with $\epsilon = 0$ was evaluated with $\epsilon = 0.25$.

Architectures	Parameters	$\epsilon = 0$		$\epsilon = 0.25$		$\epsilon = 0.5$		$\epsilon = 1$	
		Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.	Acc.	Adv. Acc.
ResNet-18	11.2M	93.20	2.84	81.86	49.02	90.30	26.96	86.14	41.99
ResNet-50	23.5M	90.03	0.17	86.10	25.52	78.73	37.67	71.34	45.59
Wide ResNet-50-2	66.9M	83.31	1.59	77.81	23.28	71.99	33.13	60.05	38.53
Wide ResNet-50-4	221.4M	82.52	1.67	78.47	22.88	70.09	31.81	59.21	38.60
ResNeXt-50 32x4d	26.5M	81.45	2.06	77.53	22.52	68.17	32.12	57.48	36.89
Densenet-161	23.0M	94.22	1.00	91.91	29.14	88.27	43.99	83.60	48.18
VGG-16	14.7M	91.20	0.02	87.88	22.72	82.38	37.67	70.20	46.63

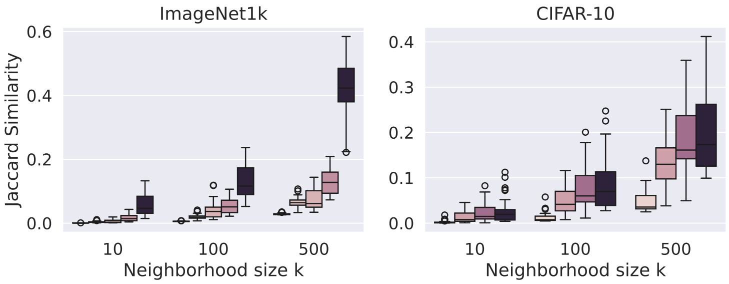

Figure 7: **Jaccard similarity with varying neighborhood size k.** Neighborhood overlap increases with larger k but trends are similar. As k increases, Jaccard Similarity becomes more similar to measures with a global perspective on similarity like CKA.

C Compute Resources

All models were trained using A100s with 80GB memory or RTX3090s (for ImageNet50 models). The training time varied depending on the dataset and model size. Training on the small CIFAR-10 dataset took around two hours at most using adversarial training. Training a robust TinyViT on ImageNet1k took around four days. Execution time for calculating model similarity was likewise dependent on the dataset as well as the measures. Reproducing the similarity results shown in this paper would take around 24 hours.

D Code and Data

Our code is available as supplementary material on OpenReview. We are in the process of uploading our model checkpoints and data to Zenodo and expect to be finished by the time the reviewers check our submission. Links to the Zenodo repositories are available in our code's README file.