BRAIN-LIKE REPRESENTATIONAL STRAIGHTENING OF NATURAL MOVIES IN ROBUST FEEDFORWARD NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representational straightening refers to a decrease in curvature of visual feature representations of a sequence of frames taken from natural movies. Prior work established straightening in neural representations of the primate primary visual cortex (V1) and perceptual straightening in human behavior as a hallmark of biological vision in contrast to artificial feedforward neural networks which did not demonstrate this phenomenon as they were not explicitly optimized to produce temporally predictable movie representations. Here, we show robustness to noise can produce representational straightening in feedforward neural networks. Both adversarial training (AT) and base classifiers for Random Smoothing (RS) induced remarkably straightened feature codes. Demonstrating their utility within the domain of natural movies, these codes could be inverted to generate intervening movie frames by linear interpolation in the feature space even though they were not trained on these trajectories. Demonstrating their biological utility, we found that RS was the best network for explaining neurons in primate V1 providing a parsimonious, bio-plausible mechanism – noise in the sensory input stages – for generating representations in early visual cortex. Finally, we compared the geometric properties of frame representations in these networks to better understand how they produced brain-like representations. Overall, this work elucidating emergent properties of robust neural networks demonstrates that it is not necessary to utilize predictive objectives or train directly on natural movie statistics to achieve models supporting more brain-like straightened movie representations that predict neural behavior.

1 INTRODUCTION

In understanding the principles underlying biological vision, a longstanding debate in computational neuroscience is whether the brain is wired to predict the incoming sensory stimulus, most notably formalized in predictive coding Rao & Ballard (1999); Friston (2009); Millidge et al. (2021), or whether neural circuitry is wired to recognize or discriminate among patterns formed on the sensory epithelium, popularly exemplified by discriminatively trained feedforward neural networksDiCarlo et al. (2012); Tacchetti et al. (2018); Kubilius et al. (2018). Arguing for a role of prediction in vision, recent work found perceptual straightening of natural movie sequences in human visual perception Hénaff et al. (2019). Such straightening is diagnostic of system whose representation could be linearly read out to perform prediction over time, and the idea of representational straightening resonates with machine learning efforts to create new types of models that achieve equivariant, linear codes for natural movie sequences. Discriminatively trained networks, however, lack any prediction over time in their supervision. It may not be surprising then that large-scale ANNs trained for classification produce representations that have almost no improvement in straightening relative to the input pixel space, while human observers clearly demonstrated perceptual straightening of natural movie sequences (subsequently also found in neurons of primary visual cortex, V1). This deficiency in standard feedforward ANNs might suggest a need for new models trained on predictive loss functions rather than pure classification to emulate biological vision.

Here, we provide evidence for an alternative viewpoint, that biologically plausible straightening can be achieved in ANNs trained for robust discrimination, without resorting to a prediction objective

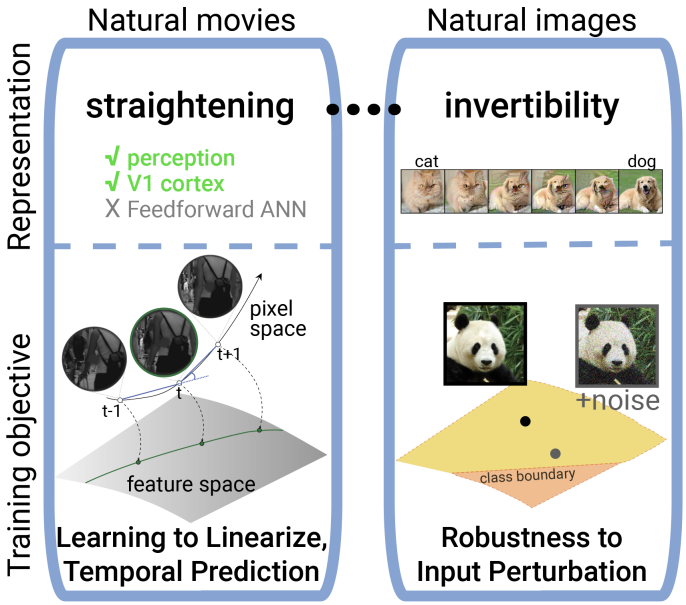

Figure 1: Perceptual straightening of movie frames can be viewed as invertibility of latent representations for static images. Left: straightening of representations refers to a decrease in the curvature of the trajectory in representation space such as neural population in the brain or human perceptual space, but standard ANNs does not show straightening Hénaff et al. (2019; 2021). Right: Invertibility of latent representation refers to interpolation between representation of two images (e.g. an image of a dog and an image of a cat), where the invertible interpolations show main features of a dog morph into main features of a cat. Invertible representations emerge in robust ANNs Engstrom et al. (2019b), obviating the need to directly train for temporal straightening.

or natural movies in training. Drawing on insights from emergent properties of adversarially-trained neural networks in producing linearly invertible latent representations, we highlight the link between perceptual straightening of natural movies to invertible latent representations learned from static images Figure1. We examine straightening in these robust feedforward ANNs finding that their properties relate to biological vision framework. The contributions of this work are as follows:

1. Robust neural networks gave rise to straightened feature representations for natural movies in their feature space, comparable to the straightening measured in the primate brain and human behavior, and completely absent from standard networks.
2. We show that linearly interpolating between the start and end frames of a movie in the output feature space of robust ANNs produces a perceptually plausible movie sequence in image space. Such invertible linear interpolation is precisely the definition of a temporally predictive feature representation.
3. Compared to prior models of early visual cortex, robustness to Gaussian input noise (Random Smoothing) is significantly better at explaining neural variance measured from V1 neurons, suggesting a new hitherto unconsidered mechanism for learning the representations in early cortical areas that achieves natural movie straightening.

2 Related Work

2.1 Mechanisms for producing brain-like representations

Feedforward ANNs as models of biological vision. Standard feedforward ANNs, although lacking a number of bio-plausible features such as feedback connections or a local learning rule Whittington & Bogacz (2019), still can explain the neural variance Schrimpf et al. (2018) recorded from rodent

Bakhtiari et al. (2021), monkeyYamins et al. (2014); Bashivan et al. (2019), and human visual cortex Khaligh-Razavi & Kriegeskorte (2014); Cichy et al. (2016) better than the alternatives which are considered more bio-plausible in terms of a prediction objective function (e.g. PredNet and CPC Zhuang et al. (2021); Schrimpf et al. (2020)). Thus, to learn the representations in the brain regardless of bio-plausibility of mechanisms, feedforward ANNs provide a parsimonious more tractable class of leading models for object recognition in the visual cortex.

Models of primary visual cortex. While standard ANNs have been the leading models for explaining neural responses in high-level visual cortex, their early layers have performed poorly at explaining early stage, V1 data. In neuroscience, rather than rely solely a on top-down training objective, there has been a tradition of explaining early visual representations using more fundamental principles such as sparse coding and predictive coding as well as invoking unsupervised training. For example, unsupervised *slow feature analysis* extracts the slow-varying features from fast-varying signals in movies based on the intuition that most external salient events (such as objects) are persistent in time, and this idea can be used to explain the emergence of complex cells in V1 Berkes & Wiskott (2005). Recent work in machine learning has attempted to blend more bottom-up principles with top-down training by experimenting with swapping out ANN early layers with V1-like models whose filters are inspired from neuroscience studies. This blended model turns out to have benefits for classification robustness in the outputs. However, it remains unclear whether there is a form of top-down training that can produce V1-like models. Such a mechanism would provide a fundamentally different alternative to prior proposals of creating a V1 through sparse coding or future prediction Hénaff et al. (2019; 2021).

2.2 Temporal prediction and invertibility in neural networks

Learning to predict over time. Changes in architecture, training diet (movies), and objective (predicting future frames) have all been explored as mechanisms to produce more explicit equivariant representations of natural movies Lotter et al. (2016); van den Oord et al. (2018). Directly related to the idea of straightening, penalizing the curvature of representations of frames was used in *Learning to linearize* Goroshin et al. (2015) to learn representations from unlabeled videos. This class of models does not need supervision which makes them more bio-plausible in nature; however, as mentioned in the previous section, they lag behind supervised feedforward ANNs both in terms of learning effective representations for object recognition and in producing brain-like representations.

Learning invertible latents. In deep learning applications, invertibility is mostly discussed in generative neural networks as a constraint to learn a prior to address the mentioned applications or to be able to reconstruct or modify real images. Usually invertibility is implemented by carefully designing dedicated architectures Jacobsen et al. (2018b); Chen et al. (2019). However, recently it has been shown it can be implemented in standard feedforward ANNs when they undergo training for adversarial robustness Engstrom et al. (2019b;c). Supporting this idea, there has been theoretical work connecting Lipschitz constraints, which resemble robustness to an l_2 ball of noise, to invertible networks. And empirical work has demonstrated that adversarially robust training encourages invertibility as linear interpolation between classes (e.g., cat to dog) results in semantically smooth image-to-image translation Engstrom et al. (2019b) as opposed to blurry image sequences produced by standard ANNs.

We reasoned that robust networks which encourage invertibility may also lead to straightening, so we sought to extend prior work and study the behavior of robustly trained networks specifically in the domain of natural movies. We report on how these networks represent natural movies in their features spaces and can invertibly reproduce movie frames in a natural sequence.

3 Methods

3.1 Baseline models

We consider the class of feedforward convolutional neural networks, typically restricting to the Resnet-50 He et al. (2015) architecture trained on ImageNet for the main analyses. Baseline networks (not trained for robustness) include a supervised Resnet-50/Resnet-101/Resnet-152, self-supervised (BarlowtwinsZbontar et al. (2021)), and Voneresnet (biological V1 front-endDapello et al. (2020)). We also consider ResNet-50 trained as a base network for action recognitionChen et al. (2021).

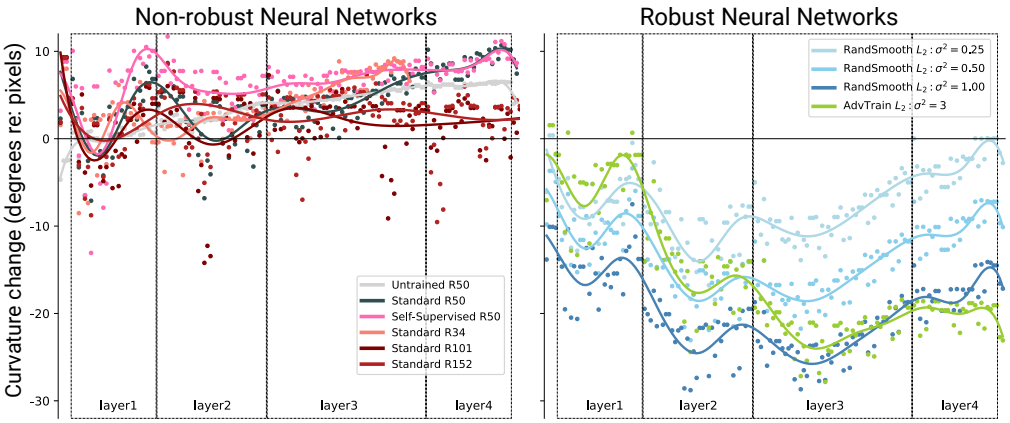

Figure 2: ANNs show straightening of representations when robustness to noise constraints (noise augmentation or robustness to adversarial attack) is added to their training. Measurements for straightening of movie sequences (from Hénaff et al. (2019), in each layer of ResNet50 architecture under different training regimes: supervised training (standard), no training (random parameters), self-supervised training Zbontar et al. (2021), supervised training with noise augmentation (base classifiers for RS) Cohen et al. (2019), and supervised training with adversarial training ($L_2, \sigma^2 = 3$) Engstrom et al. (2019a)

3.2 Models trained for robustness

We consider two forms of models trained for minimizing a classification loss $\mathcal{L}_{ce}$ in the face of input perturbations $\delta \in \mathbb{R}^{h \times w \times c}$ subject to constraints on their overall magnitude in the input space, where x, y, θ are the network input, output, and classifier parameters, respectively:

$$\mathcal{L}_{ce}(\theta, x + \delta, y) \tag{1}$$

In adversarially trained networks, projected gradient descent from the output space finds maximal directions of perturbation in the input space limited to length ϵ, and training entails minimizing the effect of these perturbation directions on the network's output Madry et al. (2018). In random smoothing Lecuyer et al. (2018); Cohen et al. (2019), a supervised network is trained but in the face of Gaussian noise added to the input space as the base classifier before performing a probabilistic inference. In this work, we only use the representations as learned in base classifiers without the probabilistic inference. The perturbations in the base classifiers δ thus can follow:

$$\delta_{rand} \sim \mathcal{N}(0, \sigma^2 I), \qquad \delta_{adv} := \underset{|\delta|_p \leq \epsilon}{\arg\max}\, \mathcal{L}_{ce}(\theta, x + \delta, y) \tag{2}$$

These defenses to input noise have different motivations. Adversarial robustness provides defense against white box attacks whereas random smoothing is protecting against general image corruptions. However, prior work has suggested a connection between corruption robustness and adversarial robustness Ford et al. (2019). Theoretically, random smoothing leads to certified robustness

Model	Clean accuracy	Robust accuracy	Model reference
RN50 standard	75.43	52.32	He et al. (2015)
RN50 AT $L_2 : \epsilon = 3$	58.50	57.81	Engstrom et al. (2019a)
RN50 AT $L_\infty : \epsilon = 4$	62.80	61.40	Engstrom et al. (2019a)
RN50 AT $L_\infty : \epsilon = 8$	48.29	47.01	Engstrom et al. (2019a)
RN50 RS $L_2 : \epsilon = 0.25$	39.40	36.01	Cohen et al. (2019)
RN50 RS $L_2 : \epsilon = 0.5$	23.75	22.21	Cohen et al. (2019)
RN50 RS $L_2 : \epsilon = 1$	10.62	10.17	Cohen et al. (2019)

Table 1: Classification accuracy of ANNs used in this work on original test images of ImageNet (clean) and on noisy images ($L_2 : \epsilon = 0.1$). Note that RS here refers to the base classifier in random smoothing without probabilistic inference.

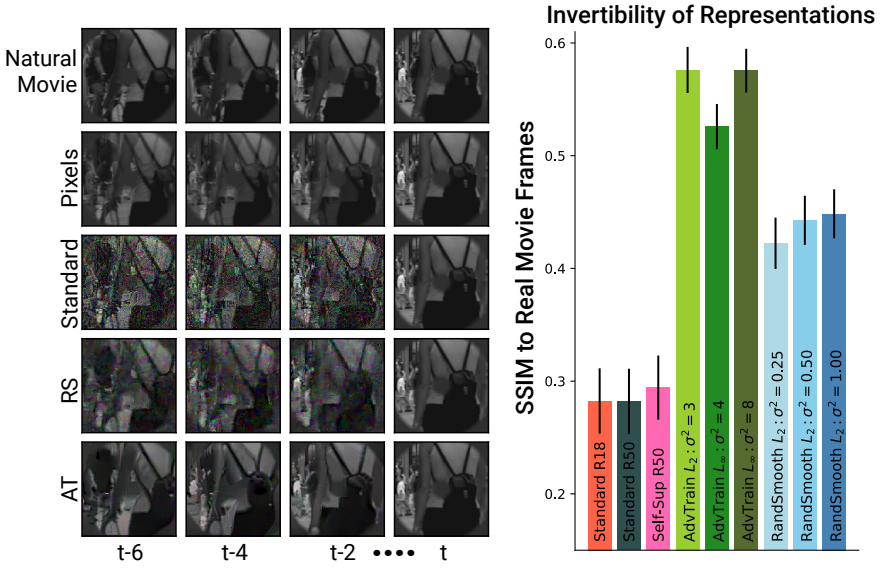

Figure 3: Invertibility as measured by the SSIM Wang et al. (2004) of the actual in-between frames (labeled as Natural Movie) and the pixel-space projected linear interpolations between the first and the last frame (labeled as Pixels) for standard ResNet50, RS (ResNet50, $L_2 : \sigma^2 = 0.5$) and AT (ResNet50, $L_2 : \sigma^2 = 3$). Interpolating the representations of the first and last frames in an invertible representation space generates a sequence of frames that are more similar to the ground-truth in-between frames, but for a non-invertible representation the generated frames are blurry and more similar to the interpolation in pixel space (a.k.a. artificial sequence). Interpolating the first and last frames in pixel space, second row, gives exactly what was called an artificial sequence in studies of straightening Hénaff et al. (2019; 2021), as opposed to natural sequence which were the actual in-between frames.

Cohen et al. (2019) and trains a condition of invertible networksJacobsen et al. (2018a), while adversarial robustness has been shown empirically to lead to invertible behavior in networks Engstrom et al. (2019b).

3.3 Representational Metrics

Straightening estimates the local curvature in a given representation r of a sequence of images (natural or artificial) of length N, $C_{seq} : \{x_{t_1}, x_{t_2}, ..., x_{t_N}\}$ as the angle between vectors connecting nearby frames, and these local estimates are averaged over the entire sequence for the overall straightening (same as Hénaff et al. (2019)):

$$c_t = cos^{-1}\left(\frac{r_t - r_{t-1}}{\|r_t - r_{t-1}\|} \cdot \frac{r_{t+1} - r_t}{\|r_{t+1} - r_t\|}\right), \qquad C_{seq} = \frac{1}{N}\sum_{t=1}^{N-1} c_t$$

Lower curvature (angle between neighboring vectors) indicates a straighter trajectory, and in the Results we generally reference curvature values to the curvature in the input space (i.e., straightening relative to pixels). This metric has been utilized in neuroscience showing that humans tend to represent nearby movie frames in a straightened manner relative to pixels (Hénaff et al. (2019)). This metric is also closely related to objectives used in efforts to train models with equivariance for natural transformations in the world as an alternative to standard networks trained for invariant object classification Goroshin et al. (2015); Sabour et al. (2017).

Expansion of a sequence of images from a movie clip is defined as the size of the minimum covering hyper-sphere in the representation r Gärtner (1999). We use this measure to supplement geometrical

characterization of a movie sequence in pixel space and in a model's representational spaces. Like curvature values, expansion values for models in the main text are referenced to the value in pixel space. We used mini-ball, a publicly available python package Gärtner (1999) to measure expansion.

4 RESULTS

4.1 ROBUST ANNS EXHIBIT REPRESENTATIONAL STRAIGHTENING

With insights from connections to invertibility (see Figure 1), we hypothesized representational straightening of movie trajectories could be present in adversarially trained robust neural networks. So we took the same movie stimuli publicly available Hénaff et al. (2019) and the same metrics, and we tested the same architecture (ResNet50 He et al. (2015)) trained under different loss functions Table 1 to perform controlled head-to-head comparisons. Figure 2 shows representational straightening of natural movies measured in layers of ResNet50 trained under AT (Engstrom et al. (2019a)) and RS (Cohen et al. (2019)) at different adversarial attack or noise levels, respectively. Robust neural networks in contrast to other ANNs decreased the curvature of natural movies. Straightening for artificial sequences (as measured in Hénaff et al. (2019)) and other models are provided in Appendix.

Perceptual Straightening measured as invertibility of latent representations. Next, we sought to empirically test how well robust networks can invert natural movies given that they contain linearized feature representation of movie frames in their high level feature spaces and given the general conceptual benefit of linearity for invertibility. We measured invertibility of each model on the same movie sequences used for measuring straightening as follows. We linearly interpolated between latent representation of the first and last frame of each movie and used the same procedure as that used previously in Engstrom et al. (2019b;a) to obtain the pixel-space correspondence of those interpolated representations. Then we used Structural Similarity index (SSIM Wang et al. (2004)) to measure how perceptually-aligned the produced inverted interpolation is to the corresponding real frame in between the first and last frames. Figure 3 shows an example of such inverted frames for standard ResNet50, RS ($L_2 : \sigma^2 = 0.5$) and AT ($L_2 : \sigma^2 = 3$), and a summary of average measured invertibility for each model. As expected, inline with the findings of previous work Engstrom et al. (2019b), AT models scored higher on invertibility of frames. However, RS models, using merely the benefits of their robustness to noisy augmentation (base classifier on top of learned representation; no probabilistic inference), also exhibit higher invertibility scores compared to standard trained models. Thus, RS models along with AT models exhibit invertibility of representations for movie frames which builds upon their ability to support perceptual straightening of natural movies in their highest layers that is similar to straightening previously measured from human subjects Hénaff et al. (2019).

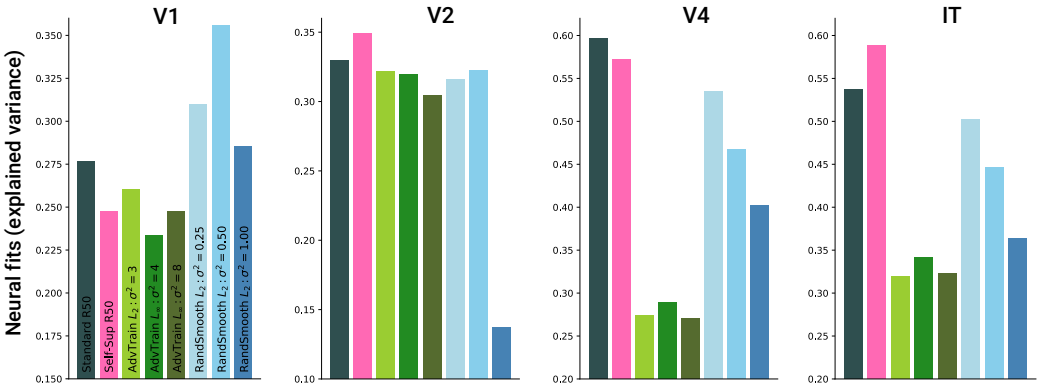

Figure 4: RS is more brain-like (V1) representations than AT. Explained variance is noise-corrected and computed as in Schrimpf et al. (2018)

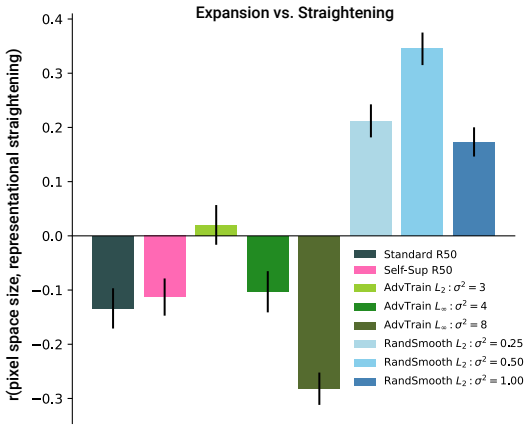

Figure 5: Can straightening for a movie sequence be explained by the size of the hyper-sphere bounding the frames (i.e. expansion in pixel space)? While RS exhibits a small but positive correlation, the rest of the models, including AT, show negative or no correlations. Positive correlation means the smaller the correlation, the more the straightened representation over the layers of the model.

4.2 Random smoothing vs. Adversarial training in explaining brain-like representations

Random smoothing as a bio-plausible mechanism underlying straightening in primary visual cortex. As shown above, straightening which is a constraint for brain-like representations in visual cortex manifests in robust neural networks. Both classes of RS and AT training for robustness to L_2 norm generate straightened representations of movie sequences. However, to distinguish among models of object recognition, we can measure how well they explain variance in patterns of activity elicited in different visual cortical areas (brain-score). Here, we measured the brain-score Schrimpf et al. (2018) using the publicly available online resource to assess the brain-likeness of each model. We found that RS models provided a better model of V1 compared to AT or non-robust models Figure4. On other benchmarks, as we go up the ventral stream hierarchy from V1 to IT, we observed a decrease in explainability of robust models, but RS models remained significantly better than AT models in higher brain areas V4 and IT. Previous work Dapello et al. (2020); Kong et al. (2022) linked adversarial robustness in models to their higher brain-score for V1, but we found that it may not be specifically driven by *adversarial* robustness, rather (L_2) noise robustness in general (as in base classifiers of RS tested here) correlates with improved explanatory power of variance in cortical area V1.

Does the size of movie frame representations in pixel space dictate straightening in downstream representations? The connection between two properties of the same representation manifold, robustness to independently sampled noise and straightened trajectories of smooth input temporal sequences, is not immediately clear. Because robustness is achieved by adding noise bounded by a norm (L_2, L_2, or L_∞) in pixel space, a natural question is whether the radius of the bounding hyper-sphere of the frames of the tested movies in pixel space (defined as expansion in Methods) was correlated with the measured straightening in feature space in each layer of the robustly trained models (Figure5). We found, however, that there seemed to be different mechanisms at play for RS versus AT in terms of achieving straightening. RS models showed (small but) positive correlations, which means the smaller the ball containing all the frames of the movie in input space, the larger the straightening effect for the representations of frames of that movie in the model. While in AT models we see the opposite (negative) or no correlation. These divergent patterns underscore differences between these models and suggest that geometric size in pixel space is not strongly constraining the degree to which a movie can be straightened.

Size of movie frame representations in feature space is relevant for capturing neural representations in V1 Between different RS models tested on different input noise levels, RS L_2 stands out as it gives a better model of V1 than those using smaller or larger magnitude input noise (Figure 4).

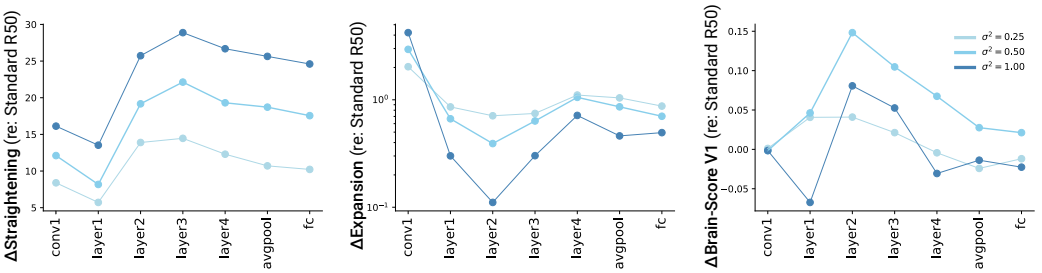

Figure 6: Geometric characteristics, straightening and curvature, of RS models related to V1 explainability. Δ means quantity is referenced to the same measure in a standard ResNet50.

For this model, we found that in addition to its intermediate level of straightening, the expansion score of movie frames in its representation was highest compared to the other RS models (Figure6, middle panel; measures are referenced to layers in a standard ResNet50 to highlight relative effect of robustness training rather than effects driven by hierarchical layer). This demonstrates a potential trade-off between improving straightening in a representation while avoiding too much added contraction of movies by robust training relative to standard training. This balance seems to be best achieved for $\sigma^2 = 0.5$, where we also see the significantly higher predictivity of V1 cortical data (Figure6, right panel).

5 Discussion

We have demonstrated novel properties of robust neural networks in how they represent natural movies. Conceptually, this work establishes a seemingly surprising connection between disparate ideas, robust discriminative networks trained on static images on one hand, to work learning to linearize by training on natural movies, on the other (Figure1). These modeling paths could both result in linearized natural movie representations. From a machine learning perspective, the invertibility and concomitant representational straightening of robust networks suggests that they learn explainable representations of natural movie statistics. Biologically, the emergence of straightening in these networks as well as their ability to best explain V1 data provides new insights into potential neural mechanisms for previously difficult to explain brain phenomena.

The leading mechanism shown here, robustness to input noise, clearly outperformed robustness to engineered noise in biological predictions. So biological constraints could lend parsimony to selecting among models, each with a different engineering goal. RS gains traction as a more simple, and powerful way of achieving robustness in ANNs, which is inline with a long history of probabilistic inference in visual cortex of humans Pouget et al. (2013). In our study of straightening in RS we saw the deeper the layer, the less straightened as opposed to AT, which suggests maybe internal neurons deep in visual cortex, e.g. IT neurons may not be quite as robust as the later layers of AT models as confirmed in a recent study Guo et al. (2022). Indeed, the best fitting model of V1 was not the most robust nor the most straightened but tended to also show the least contracted representations – consistent with a known dimensionality expansion from the sensory periphery to V1 in the brain. Future work exploring a wider variety of robustness training in conjunction with more bioplausible architectures, objectives, and training diets may yet elucidate the balance of factors contributing to biological vision.

REFERENCES

Shahab Bakhtiari, Patrick J Mineault, Tim Lillicrap, Christopher C Pack, and Blake Aaron Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=t1czgrQOrwW.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, May 2019.

P Berkes and L Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vis.*, 5(6):9–9, July 2005.

Chun-Fu Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

Ricky T. Q. Chen, Jens Behrmann, David K Duvenaud, and Joern-Henrik Jacobsen. Residual flows for invertible generative modeling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/5d0d5594d24f0f955548f0fc0ff83d10-Paper.pdf.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. March 2020.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.*, 6(1), September 2016.

Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing, 2019. URL https://arxiv.org/abs/1902.02918.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D. Cox, and James J. DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *bioRxiv*, 2020. doi: 10.1101/2020.06.16.154542. URL https://www.biorxiv.org/content/early/2020/10/22/2020.06.16.154542.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, February 2012.

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

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. June 2019b.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Learning perceptually-aligned representations via adversarial robustness. In *ArXiv preprint arXiv:1906.00945*, 2019c.

Nic Ford, Justin Gilmer, Nicolas Carlini, and Dogus Cubuk. Adversarial examples are a natural consequence of test error in noise. January 2019.

Karl Friston. The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.*, 13(7): 293–301, July 2009.

Bernd Gärtner. Fast and robust smallest enclosing balls. In Jaroslav Nešetřil (ed.), *Algorithms - ESA' 99*, pp. 325–338, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg. ISBN 978-3-540-48481-3.

Ross Goroshin, Michael F Mathieu, and Yann LeCun. Learning to linearize under uncertainty. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/eefc9e10ebdc4a2333b42b2dbb8f27b6-Paper.pdf.

Chong Guo, Michael J Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, and James J DiCarlo. Adversarially trained neural representations may already be as robust as corresponding biological neural representations. June 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Olivier J Hénaff, Robbe L T Goris, and Eero P Simoncelli. Perceptual straightening of natural videos. *Nat. Neurosci.*, 22(6):984–991, June 2019.

Olivier J Hénaff, Yoon Bai, Julie A Charlton, Ian Nauhaus, Eero P Simoncelli, and Robbe L T Goris. Primary visual cortex straightens natural video trajectories. *Nat. Commun.*, 12(1):5982, October 2021.

Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. November 2018a.

Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-RevNet: Deep invertible networks. February 2018b.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, 10(11):e1003915, November 2014.

Nathan C L Kong, Eshed Margalit, Justin L Gardner, and Anthony M Norcia. Increasing neural network robustness improves match to macaque V1 eigenspectrum, spatial frequency preference and predictivity. *PLoS Comput. Biol.*, 18(1):e1009739, January 2022.

Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L K Yamins, and James J DiCarlo. CORnet: Modeling the neural mechanisms of core object recognition. September 2018.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. February 2018.

William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. May 2016.

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*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Beren Millidge, Anil Seth, and Christopher L Buckley. Predictive coding: A theoretical and experimental review. July 2021.

Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.*, 16(9):1170–1178, September 2013.

R P Rao and D H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.*, 2(1):79–87, January 1999.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. October 2017.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018. URL https://www.biorxiv.org/content/10.1101/407007v2.

Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020. URL https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X.

Andrea Tacchetti, Leyla Isik, and Tomaso A Poggio. Invariant recognition shapes neural representations of visual input. *Annu. Rev. Vis. Sci.*, 4:403–422, September 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. July 2018.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

James C.R. Whittington and Rafal Bogacz. Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, 23(3):235–250, 2019. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2018.12.005. URL https://www.sciencedirect.com/science/article/pii/S1364661319300129.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111. URL https://www.pnas.org/doi/abs/10.1073/pnas.1403112111.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. March 2021.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021. doi: 10.1073/pnas.2014196118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2014196118.

A APPENDIX

A.1 STRAIGHTENING FOR BOTH NATURAL AND ARTIFICIAL SEQUENCES

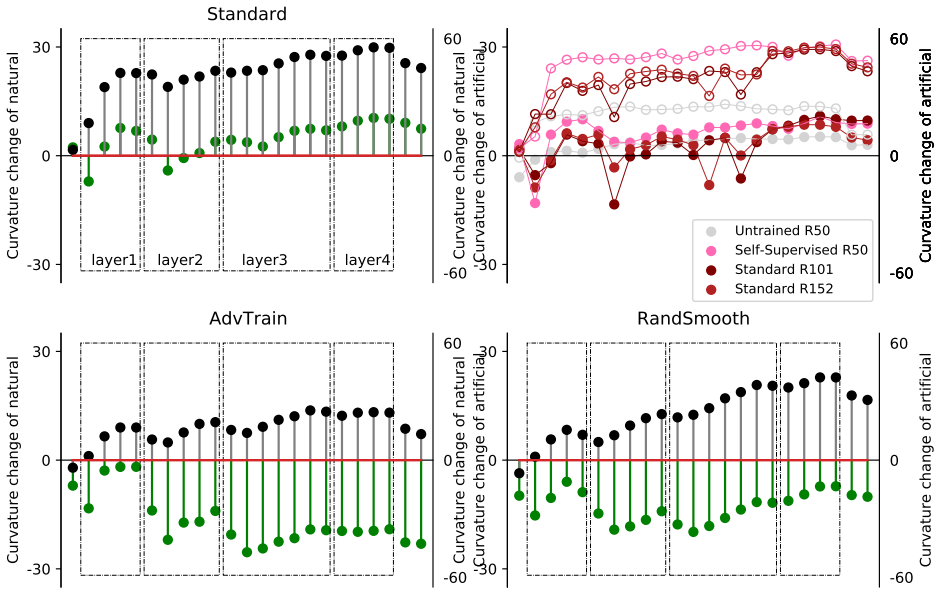

Figure 7: ANNs show straightening of representations when robustness to noise constraints (noise augmentation or robustness to adversarial attack) is added to their training. From left to right, measurements for straightening of movie sequences (from Hénaff et al. (2019), natural sequence: green, artificial sequence: gray) in each layer of ResNet50 architecture under different training regimes: supervised training (standard), no training (random parameters), self-supervised training Chen et al. (2020), supervised training with noise augmentation ($L_2, \sigma^2 = 0.5$) Cohen et al. (2019) and supervised training with adversarial training ($L_2, \sigma^2 = 3$) Engstrom et al. (2019a)

A.2 STRAIGHTENING IN OTHER ARCHITECTURES

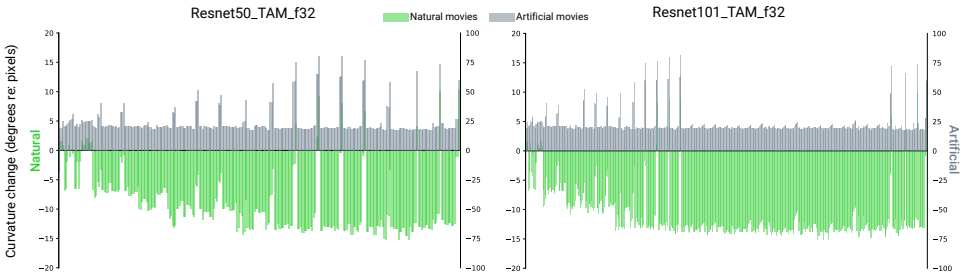

Figure 8: Straightening for ResNet50 and ResNet101 trained as the base architecture for action recognition Chen et al. (2021). Although these models were not trained for straightening, they exhibit small but measurable straightening for natural movies. However, the curvatures for artificial sequences were not increased as much as curvature increase for artificial sequences in robust neural networks 7.

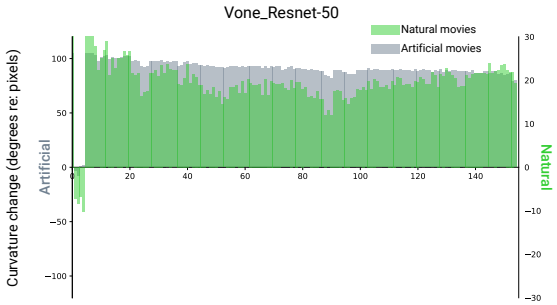

Figure 9: Lack of straightening for natural movies in ResNet50 trained with a biologically-inspired model of V1 in the front-end. Dapello et al. (2020) Vone-Resnet50 exhibit robustness to adversarial attack but the fact that it does not exhibit straightening (except for the front-end) provides yet another evidence that adversarial robustness does not always accompanied straightening.

A.3 BRAIN-SCORE FOR ALL MODELS ACROSS ALL LAYERS

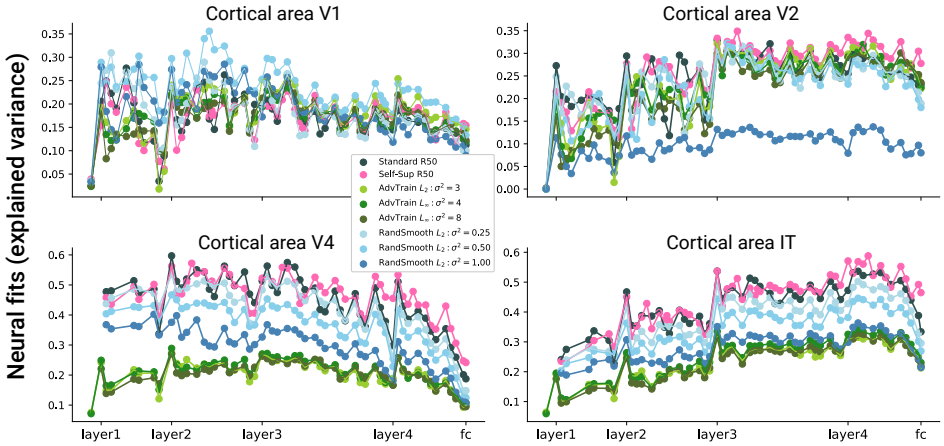

Figure 10: Brain-score (data and metric publicly available Schrimpf et al. (2018)) for the models used in this study. The bar plot Figure 4is a summary of this plot.

A.4 GEOMETRICAL CHARACTERIZATION OF AT NETWORKS RELATES TO BRAIN-LIKENESS

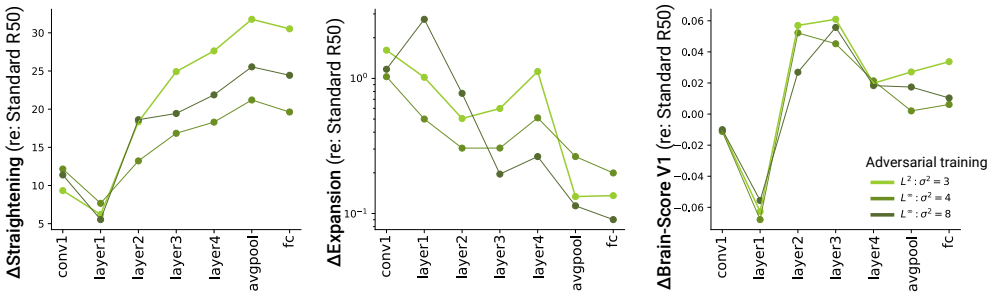

Figure 11: Geometric characteristics, straightening and curvature, of AT models and their relation to V1 explainability. Δ means quantity is referenced to the same measure in a standard ResNet50.