Are models trained on temporally-continuous data streams more adversarially robust?

Anonymous Author(s)
Affiliation
Address
email

Abstract

Task-optimized convolutional neural networks are the most quantitatively accurate models of the primate visual system. Unlike humans, however, these models can easily be fooled by modifying their inputs with human-imperceptible image perturbations, resulting in poor *adversarial robustness*. Prior work showed that modifying a model's training objective or its architecture can improve its adversarial robustness. Another ingredient in building computational models of sensory cortex is the training dataset and, to our knowledge, its effect on a model's adversarial robustness has not be investigated. Motivated by observations that chicks develop more invariant visual representations with more temporally-continuous visual experience, we here evaluate a model's adversarial robustness when it is trained on a more naturalistic dataset—a longitudinal video dataset collected from the perspective of infants (SAYCam; Sullivan et al., 2020). By evaluating the adversarial robustness of models on 26-way classification of a set of annotated video frames from this dataset, we find that models that have been pre-trained on SAYCam video frames are more robust than those that have been pre-trained on ImageNet. Our results suggest that to build models that are adversarially robust, additional efforts should be made in curating datasets that are more similar to the natural image sequences and the visual experience infants receive.

1 Introduction

Task-optimized deep convolutional neural networks (CNNs) are the most quantitatively accurate models of the human and non-human primate ventral visual stream (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018). However, their image-classification behavior on images that have been modified by non-random, human-imperceptible image perturbations diverges significantly from that of humans. Specifically, these *adversarial perturbations* can cause CNNs to confidently misclassify an image even though they can classify the unperturbed image correctly (Szegedy et al., 2014). We view the problem of building computational models (of the visual system) that are robust to these perturbations through the lens of the goal-driven modeling approach (Yamins & DiCarlo, 2016), which consists of four major components: objective function, model architecture, data stream, and learning rule. Each component can be tweaked in order to better achieve an optimization goal, which, in this work, is adversarial robustness.

Some work focused on reformulating a generic objective function such as the cross-entropy loss in image classification to improve the adversarial robustness of CNNs (e.g., Madry et al., 2018; Ross & Doshi-Velez, 2018; Zhang et al., 2019). Other work focused on architectural modifications to improve adversarial robustness (e.g., Vuyyuru et al., 2020; Dapello et al., 2020). To our knowledge, however, the impact of a model's *training data* on its adversarial robustness has not been investigated before and therefore is the focus of the work presented here.

Submitted to 3rd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM 2021) of the Neural Information Processing Systems (NeurIPS) conference.

We were motivated to study the effects of the data stream on robustness because of the observations by Wood & Wood (2016) and by Wood et al. (2016). Specifically, they found that newborn chicks reared with more temporally continuous visual experience were better able to generalize to novel viewpoints of objects and also had better object recognition abilities than chicks reared with less temporally continuous visual experience. As the visual experience acquired via training on ImageNet (Deng et al., 2009) is highly unrealistic (developmentally) relative to that received by human infants due to, for example, the sheer number of category labels or the lack of temporal continuity in the data, we reasoned that training models on a more developmentally-realistic data distribution (SAYCam; Sullivan et al., 2020) could lead to visual representations that are more robust to input perturbations.

Our work is closely related to that of Orhan et al. (2020) and of Zhuang et al. (2021) who both used the same infant head-cam video dataset (SAYCam) for training models. Orhan et al. (2020) showed that models can learn visual representations useful for basic downstream tasks by training them on head-cam video frames using self-supervised techniques. Similarly, Zhuang et al. (2021) showed that models trained using the head-cam dataset are also competitive models of the ventral visual stream. In both cases, however, the model's adversarial robustness was not evaluated.

2 Methods

Dataset We used video frames extracted from a subset of the SAYCam infant head-cam dataset (Sullivan et al., 2020), which is a longitudinal, egocentric dataset of videos captured using head cameras placed on three infants. Specifically, we used the videos from child S and child Y and divided them into approximately five-minute episodes. Video frames were then extracted from these episodes at five frames per second, as in Orhan et al. (2020). In order to evaluate models trained using these images on a downstream classification task and on adversarial robustness, we used the annotated video frames from child S provided by Orhan et al. (2020). In the annotated dataset, there are approximately $58\,000$ annotated video frames where 50% of those belonged to the train set and the remaining 50% belonged to the validation set. Each of these images belonged to one of 26 categories, including car, cat, and plush animal.

Model architecture We fixed the model architecture to be ResNet-18 (He et al., 2016), as it has been shown to be one of the best models of the ventral visual stream when trained on supervised image classification (Schrimpf et al., 2018). This architecture has also been shown to be a highly accurate model of the ventral visual stream when trained in a self-supervised manner (Zhuang et al., 2021).

Objective functions For models trained using ImageNet, two objective functions were used: cross-entropy loss for supervised learning (from PyTorch's model zoo) and simple siamese representation learning for self-supervised learning (SimSiam; Chen & He, 2021). SimSiam is a state-of-the-art contrastive learning algorithm that does not need a memory bank or large batches during training. For models trained using video frames from SAYCam, two self-supervised objective functions were used: SimSiam (as above) and temporal classification. The temporal classification objective encourages models to be able to classify video frames according to the five-minute episode to which they belong (Orhan et al., 2020).

Model training For the temporal classification objective, we used similar hyperparameters and the same image augmentation scheme as in the code base of Orhan et al. (2020). Image augmentations used when training models on SAYCam using SimSiam are the same as those used by Chen & He (2021). Additional model training and image augmentation details can be found in Appendix A.2.

Linear classifier training Models trained using each of the dataset-objective-function combinations were evaluated on their 26-way linear classification performance on the set of annotated video frames from child S, similar to the procedure of Orhan et al. (2020). If the model had been trained in a supervised manner (e.g., supervised ImageNet), its final fully-connected (`fc`) layer was removed and replaced with a linear readout of dimensions 512×26 (i.e., the linear readout was placed post-average pool in ResNet-18). For self-supervised models, an `fc` layer of the same dimensions as above was appended to the end of those models which do not have `fc` layers. Additional model training and image augmentation details can be found in Appendix A.2.

Adversarial robustness evaluation Adversarial robustness was evaluated on the 26-way linear classification task so that each of our models pre-trained (in a supervised or a self-supervised manner) on either ImageNet or SAYCam had a linear readout layer for the 26-way class predictions. We focused primarily on "white-box" adversarial attacks and used projected gradient descent (PGD; Madry et al., 2018) to find the image perturbations that maximized the cross-entropy loss (i.e., untargeted attacks). Adversarial examples were generated using 20 steps of PGD with step sizes of $\varepsilon \times 2/20$, where ε is the maximum allowed "size" of the perturbation. The sizes of the perturbations were measured using the ℓ_∞-, ℓ_2-, and ℓ_1-norm. Additional details on the size constraints can be found in Appendix A.3.

3 Results

3.1 Adversarial Examples

Using a ResNet-18 pre-trained in a supervised manner on ImageNet along with its linear readout trained to perform the 26-way classification, we generated a few example adversarial images for demonstration, shown in Figure 3 (found in Appendix A.1). The adversarial images were generated using perturbations constrained by the ℓ_∞-norm and $\varepsilon = 1/1020$ (i.e., $\|\boldsymbol{\delta}\|_\infty \leq 1/1020$, where $\boldsymbol{\delta}$ is the perturbation). This provides additional evidence that human-imperceptible image perturbations can drastically affect a model's behavior.

3.2 Adversarial Accuracy

Using models trained on either ImageNet or the SAYCam dataset, we trained 26-way linear readouts on top of the (frozen) pre-trained backbones to perform classification of annotated video frames from child S. Here we directly compare the clean and the adversarial accuracy on the 26-way classification. Table 1 shows the accuracy on the classification task when the images have been adversarially perturbed by relatively weak constraints. We found that models trained using ImageNet have comparable accuracy on unperturbed images (Table 1; Clean Accuracy), but have worse accuracy on images that have been weakly adversarially perturbed. When comparing "SimSiam ImageNet" with "SimSiam Y", where *only* the training dataset is different (same architecture and objective), we find that even with comparable clean accuracy, "SimSiam Y" exhibits higher adversarial accuracy for all perturbation types.

Across all the tested perturbation sizes, we observed that models trained on either child S's or child Y's video frames exhibited higher adversarial robustness than those trained on ImageNet, shown in Figure 1. As expected for each norm constraint, model performance on the linear classification task plummets to zero as the maximum allowable perturbation size increases. However, the model pre-trained in a supervised manner on ImageNet performed the worst overall, with its accuracy dropping off the fastest (Figure 1; blue).

Table 1: **Clean and adversarial accuracy for each model trained on each dataset-objective-function combination on the 26-way classification of the annotated video frames from child S.** Each column is associated with a different type of constraint applied on the perturbation. S and Y in the model names refer to which dataset from SAYCam the model was trained on. For example, "SimSiam Y" means that the model was trained on child Y videos using the SimSiam objective. The performance values reported here are the accuracy across the given size constraint for each norm. For the "Random" model, a 26-way linear readout was trained on top of a randomly initialized ResNet-18.

Model	**Clean Accuracy**	$\|\boldsymbol{\delta}\|_\infty \leq 1/1020$	$\|\boldsymbol{\delta}\|_2 \leq 0.15$	$\|\boldsymbol{\delta}\|_1 \leq 40$
Supervised ImageNet	55.8%	2.7%	8.7%	8.1%
SimSiam ImageNet	46.6%	12.5%	18.5%	17.5%
SimSiam S	37.9%	16.4%	21.0%	19.8%
SimSiam Y	46.7%	20.9%	27.7%	26.3%
Temporal Classification S	63.1%	22.5%	32.2%	30.1%
Temporal Classification Y	54.4%	19.4%	27.6%	25.7%
Random	28.3%	11.3%	17.7%	18.4%

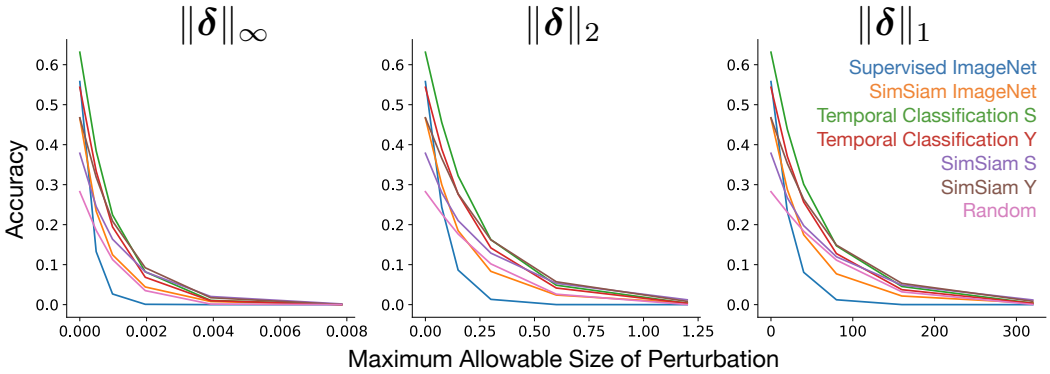

Figure 1: **Adversarial accuracy on the 26-way classification of annotated frames from child S as a function of perturbation size for each norm.** As the perturbation size increase, we see that the accuracy of all the models is driven to zero. We refer the reader to Appendix A.3 for further details on the size constraints used. At the low end of the perturbation size regime, we find that models pre-trained on the head-cam dataset (either using videos from child S or child Y; green, red, purple, and brown) have higher adversarial accuracy compared to that of models pre-trained on ImageNet (blue and orange). We do not show the full range of the perturbation sizes used in the evaluation—the smallest six of eight sizes are plotted here—because the accuracy of all the models drops to zero for the largest sizes.

3.3 Spatial Frequency Tuning Analysis

We next asked whether there were differences in spatial frequency tuning preferences between models that were trained on ImageNet and those that were trained on SAYCam. We investigated this property because prior work has suggested that the brittleness of these models can be associated with their sensitivity to image features of high spatial frequencies (Yin et al., 2019; Wang et al., 2020; Kong et al., 2021). To generate spatial frequency tuning curves for each artificial neuron, we presented each model with a set of fixed size Gabor patches that varied in spatial frequency, phase, and orientation and "recorded" activations for each neuron in the `maxpool` layer of ResNet-18. Each tuning curve (for each neuron) was then summarized by the spatial frequency at which the tuning curve reached its maximum value, resulting in a distribution of preferred spatial frequencies for the neurons in the `maxpool` layer.

The preferred spatial frequency distributions for the models trained on ImageNet and on SAYCam are shown in Figure 2. Consistent with other findings, we found that a large proportion of artificial neurons in less adversarially robust models were tuned to high spatial frequencies (Figure 2; blue and orange). In contrast, models that were more adversarially robust had a relatively smaller proportion of neurons tuned to the highest spatial frequency bin (Figure 2; green, red, purple, and brown) and more neurons were tuned to the middle spatial frequencies.

4 Discussion

Current state-of-the-art computational models of the primate visual system are task-optimized deep CNNs. A closer look at these models, however, reveals fundamental ways in which human and machine perception differ. In particular, these models are notoriously brittle, where human-imperceptible image perturbations can drastically change their behavior, resulting in clear safety issues in real-world applications. Inspired by the work of Wood & Wood (2016) and of Wood et al. (2016), we hypothesized that a more temporally-continuous (i.e., natural) data stream could lead to improved visual representations. In this work, we provided an initial investigation into how the training data distribution can affect a model's adversarial robustness. We found that pre-training models using infant head-cam videos resulted in higher adversarial robustness than that of models pre-trained using ImageNet and that there were a relatively smaller proportion of artificial neurons in those models tuned to the highest spatial frequencies.

We first note that our results cannot distinguish between whether it is the "statistics" of the modality of video frames (e.g., higher temporal correlations) or the "statistics" of the infant visual experience

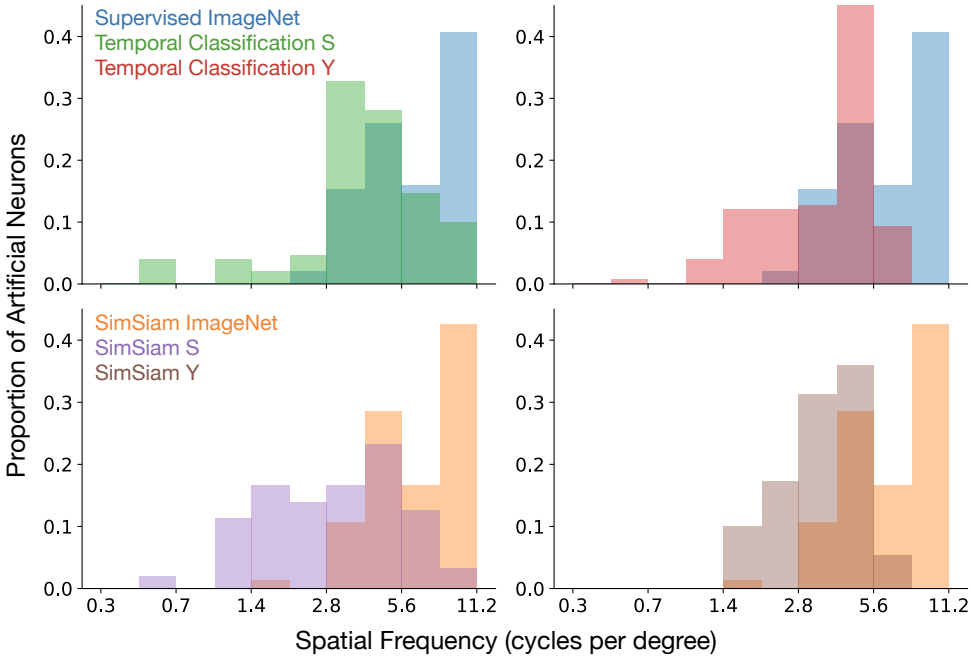

Figure 2: **Preferred spatial frequency distribution of artificial neurons in models trained on different datasets and objective functions.** The preferred spatial frequency distribution of neurons in the `maxpool` layer of these models shows that models pre-trained on ImageNet (blue, orange) contain a substantial proportion of neurons tuned to high spatial frequencies. This is in contrast to models trained on SAYCam (green, red, purple, and brown), where a larger proportion of neurons are tuned to the middle spatial frequencies.

per se that are the drivers of improved adversarial robustness. Therefore, we do not claim that infant head-cam video datasets are necessary to build models that are more adversarially robust. We also note that the adversarial attacks performed here are relatively weak, as the maximum allowable perturbation sizes presented in Table 1 are relatively small. More work is needed to boost the robustness of these models in larger perturbation size regimes.

We suggest a few possible future directions for this work. First, we fixed the architecture to be ResNet-18 with 2D convolutions. Thus, we did not make use of the temporally continuous nature of SAYCam architecturally (the temporal classification objective encourages models to group frames belonging to the same time window). Could architectures that explicitly take advantage of the temporal continuity in video datasets learn visual representations that are more robust to perturbations? For example, 3D-ResNets have been able to achieve success in action recognition tasks (Hara et al., 2018). Can spatiotemporal filters contribute to adversarial robustness? More generally, can training on video streams improve adversarial robustness?

On the data stream front, we investigated the impact of training models on more temporally-continuous (and developmentally-realistic) inputs on adversarial robustness. This, however, can be made even more biologically plausible. In particular, infant visual acuity is known to be almost an order of magnitude lower than that of an adult (Mayer & Dobson, 1982; Norcia & Tyler, 1985). Thus, visual cortex effectively receives low-resolution inputs early in development. Could training models on blurred inputs, in concert with inputs that infants are more likely to receive, benefit adversarial robustness?

Finally, we note that in training these models, the entire dataset—video frames from six months to 30 months of age for child S—is "replayed" over and over again (once every training epoch). This is clearly not the visual experience infants receive. Building models that are able to learn visual representations in an "online" manner would be necessary for them to be more biologically plausible and can perhaps provide insight about other parts of the brain that may be recruited to facilitate learning.

5 Acknowledgements

Having a large enough training dataset is one of the most important, if not *the* most important, components of building models that can mimic human capabilities and further our understanding of sensory cortex. We are therefore grateful to Sullivan et al. (2020) for their Herculean efforts in collecting and in making available the longitudinal infant head-cam video dataset. We are also grateful to A. Emin Orhan for providing initial assistance with the SAYCam dataset.

References

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. In *Advances in Neural Information Processing Systems*, pp. 13073–13087, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.

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*, pp. 770–778, 2016.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nathan CL 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. *bioRxiv*, 2021.

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.

D Luisa Mayer and Velma Dobson. Visual acuity development in infants and young children, as assessed by operant preferential looking. *Vision Research*, 22(9):1141–1151, 1982.

Anthony M Norcia and Christopher W Tyler. Spatial frequency sweep vep: visual acuity during the first year of life. *Vision Research*, 25(10):1399–1408, 1985.

Emin Orhan, Vaibhav Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9960–9971, 2020.

Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2018.

Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C Frank. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, pp. 1–10, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Manish Reddy Vuyyuru, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2135–2146, 2020.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8684–8694, 2020.

Justin N Wood and Samantha MW Wood. The development of newborn object recognition in fast and slow visual worlds. *Proceedings of the Royal Society B: Biological Sciences*, 283(1829):20160166, 2016.

Justin N Wood, Aditya Prasad, Jason G Goldman, and Samantha MW Wood. Enhanced learning of natural visual sequences in newborn chicks. *Animal Cognition*, 19(4):835–845, 2016.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.

Daniel LK 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.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), 2021.

A Appendix

A.1 Some adversarial examples

Figure 3 shows two adversarial examples generated using a model pre-trained on ImageNet in a supervised manner.

A.2 Additional model training details

Here we report additional details of model training and provide the image augmentations used via APIs belonging to the `torchvision.transforms` package.

Temporal classification Models trained on temporal classification were trained for 20 epochs using the Adam optimizer (Kingma & Ba, 2014), a batch size of 512, weight decay of 1×10^{-9}, and an initial learning rate of 1×10^{-3} that was decayed by a factor of 10 at epochs 10 and 15. Two GPUs were used to train each of these models. The following image augmentations were used:

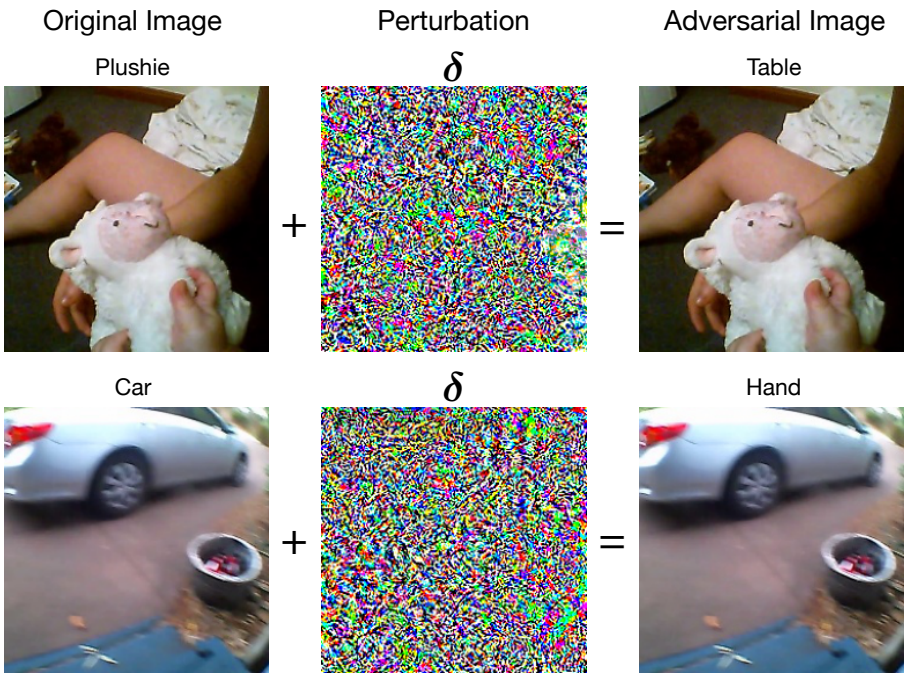

Figure 3: **Human-imperceptible image perturbations can drastically affect model predictions.** The left column shows the original images along with their correct model predictions. The right column shows their corresponding adversarial images and their incorrect model predictions. The middle column shows the perturbations (with $\|\boldsymbol{\delta}\|_\infty \leq 1/1020$) that were added to the original images to obtain the adversarial images. Note that their values have been magnified so that the perturbations are visible. The adversarial images were generated from the annotated video frames of child S using a model pre-trained on ImageNet in a supervised manner.

```
transforms.Compose([
    transforms.RandomResizedCrop(224, scale=(0.2, 1.)),
    transforms.RandomApply(
        [transforms.ColorJitter(0.9, 0.9, 0.9, 0.5)],
        p=0.9
    ),
    transforms.RandomGrayscale(p=0.2),
    transforms.RandomApply(
        [GaussianBlur(sigma_min=0.1, sigma_max=2.0)],
        p=0.5
    ),
    transforms.RandomHorizontalFlip(0.5),
    transforms.ToTensor(),
    transforms.Normalize(
        mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]
    ),
])
```

SimSiam Models that were trained using SimSiam and the SAYCam dataset were optimized for 20 epochs using stochastic gradient descent (Bottou, 2010) with momentum of 0.9, a batch size of 256, weight decay of 1×10^{-9}, and an initial learning rate of 0.1 that was updated using a cosine annealing schedule. Two GPUs were used to train each of these models. The model that was trained using SimSiam and ImageNet was optimized for 100 epochs using stochastic gradient descent with momentum of 0.9, a batch size of 512, weight decay of 1×10^{-4}, and an initial learning rate that was updated using a cosine annealing schedule. Four GPUs were used for training this model. The following image augmentations were used (identical to those of Chen & He, 2021):

```
transforms.Compose([
    transforms.RandomResizedCrop(224, scale=(0.2, 1.0)),
    transforms.RandomHorizontalFlip(),
    transforms.RandomApply(
        [
            transforms.ColorJitter(
                brightness=0.4, contrast=0.4, saturation=0.4, hue=0.1
            )
        ],
        p=0.8,
    ),
    transforms.RandomGrayscale(p=0.2),
    transforms.RandomApply(
        [GaussianBlur(sigma_min=0.1, sigma_max=2.0)],
        p=0.5
    ),
    transforms.ToTensor(),
    transforms.Normalize(
        mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]
    ),
])
```

Linear classifier training In most cases, the linear readout weights were trained for 100 epochs to minimize the cross-entropy loss using the Adam optimizer, a batch size of 256, weight decay of 0.0, and a constant learning rate of 5×10^{-4}. The "Random" model was trained for 100 epochs with an initial learning rate of 1×10^{-3} that was decayed by a factor of 10 at epoch 75. For most models that were pre-trained using the SimSiam objective, the linear readout weights were trained for 200 epochs with a constant learning rate of 5×10^{-4}. The "SimSiam S" model was trained with an initial learning rate of 1×10^{-3} that was decayed by a factor of 10 at epoch 120. A single GPU was used to train the linear classifier for all models. The following image augmentations were used:

```
transforms.Compose([
    transforms.RandomHorizontalFlip(0.5),
    transforms.ToTensor(),
    transforms.Normalize(
        mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]
    ),
])
```

A.3 Adversarial perturbation size constraints

Here we provide additional details on the size constraints applied to the adversarial perturbations, $\boldsymbol{\delta}$. Note that the same size constraints were used by Dapello et al. (2020).

- ℓ_∞-norm: $\|\boldsymbol{\delta}\|_\infty \leq \varepsilon, \varepsilon \in \{0, 1/2040, 1/1020, 1/510, 1/255, 2/255, 4/255, 8/255\}$
- ℓ_2-norm: $\|\boldsymbol{\delta}\|_2 \leq \varepsilon, \varepsilon \in \{0, 0.075, 0.15, 0.3, 0.6, 1.2, 2.4, 4.8\}$
- ℓ_1-norm: $\|\boldsymbol{\delta}\|_1 \leq \varepsilon, \varepsilon \in \{0, 20, 40, 80, 160, 320, 640, 1280\}$