Out-of-distribution robustness: Limited image exposure of a four-year-old is enough to outperform ResNet-50

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent gains in model robustness towards out-of-distribution images are predominantly achieved through ever-increasing large-scale datasets. While this approach is very effective in achieving human-level distortion robustness, it raises the question of whether human robustness, too, requires massive amounts of experience. We therefore investigated the developmental trajectory of human object recognition robustness by comparing children aged 4–6, 7–9, 10–12, 13–15 against adults and against different neural networks. Assessing how recognition accuracy degrades when images are distorted by salt-and-pepper noise, we find that while overall performance improves with age, even the youngest children in the study showed remarkable robustness and outperformed standard CNNs on moderately distorted images.

Using a back-of-the-envelope calculation, we then estimated the number of 'images' that those young children had been exposed to during their lifetime. Even if we assume that a new image is seen every 2 seconds of wake time, children aged 4–6 only saw approximately 50M images, which is already lower than 90 epochs $\times$ 1.3M images during standard ImageNet training. This indicates that human out-of-distribution robustness develops very early and may not require seeing billions of different images during lifetime given the right choice of representation and information processing optimised during evolution.

1 Introduction

Not too long ago, humans outperformed convolutional neural networks (CNNs) on out-of-distribution (OOD) datasets [e.g., 1–5]. However, recently this longstanding robustness gap between humans and CNNs is beginning to close. Currently, the best state-of-the-art (SOTA) models are matching, or even exceeding human performance on most OOD datasets [6]. Even though some of these models have an innovative architecture and/or training procedure—such as CLIP [7] or other vision transformers [8]—the most crucial feature to achieve human-like OOD robustness appears to be *large-scale training*. While standard training on ImageNet includes 1.3M images, most models showing human-like OOD robustness are trained on much larger datasets—ranging from 14M (Big Transfer models [9]) to staggering 940M images (semi-weakly supervised models [10]). While both architecture and data matter (e.g., vision transformers trained on ImageNet are better than standard CNNs trained on ImageNet); even standard CNNs trained on large-scale datasets (such as Big Transfer models) achieve remarkable robustness. This indicates that large-scale training may be *sufficient* for OOD robustness. Here we ask whether large-scale training is also *necessary* to achieve OOD robustness. If so, we would expect human OOD robustness to be low in early childhood, and to

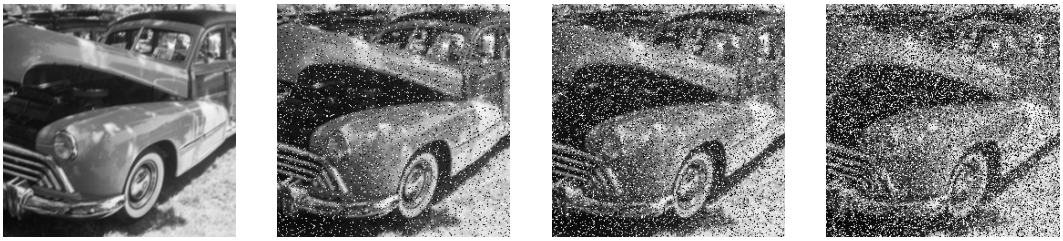
Submitted to 3rd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM 2021) of the Neural Information Processing Systems (NeurIPS) conference.

increase with age (= more input) (hypothesis I). Alternatively, human OOD robustness might instead result from clever information processing and representation that copes with less data. In this case, we would expect human OOD robustness to be already high in early childhood (hypothesis II).

Both hypotheses can be evaluated with respect to developmental data. However, behavioral research investigating the development of object recognition (robustness) in children (after 2 years of age) and adolescents is extremely sparse and several reviews have pointed out the lack of such studies [11–13]. Here, we present a detailed investigation of the developmental trajectory of object recognition robustness. We gathered 7,200 psychophysical trials from children, adolescents and adults and compared their performance with two models; one SOTA model trained on a large-scale dataset (CLIP) and a standard CNN trained on ImageNet (ResNet-50). We then looked at children's OOD robustness as a function of the estimated number of input images (i.e., dataset-size) at different time points during development. In doing so, we investigated whether human OOD robustness develops early during development and if—similarly to SOTA models—human OOD robustness is dependent on large-scale visual input.

2 Methods

The methods used in this study are adapted from a series of psychophysical experiments conducted in the Wichmann-lab [2]. Some aspects were slightly modified in order to get a paradigm suitable to test children (see Appendix A.1 for more information). To obtain an OOD dataset, we added salt-and-pepper noise to 320 ImageNet-images (see Appendix A.2 for more details on the dataset). Figure 1 shows some example stimuli. In total, we collected 7,200 trials from a sample of 46 children/adolescents (4–15 years) and three adults. The study was approved by the institutional ethical review board. All participants reported normal or corrected to normal vision and provided (parental) written consent. In each trial, participants were presented with an image and had to to choose the corresponding category out of 16 basic entry-level categories (such as cat, knife, truck, etc.). Stimuli were balanced with regard to those categories and presented in a pseudo-random order (see Appendix A.2 for more information). To compare humans and models we mapped the 1,000 class decision vector of ImageNet-trained models to the 16 classes using the WordNet hierarchy [14]. Stimuli (256×256 pixels) were presented at the centre of a 14" screen with a spatial resolution of 1920×1200 pixels at a refresh rate of 120 Hz. Given a viewing distance of 60 cm, retinal image size was 4×4 degrees of visual angle. Stimuli were then followed by a pink noise mask with $1/f$ spectral shape. Short presentation time and noise mask was employed to ensure fair comparison between humans and feedforward neural networks. We evaluated two different models using the same stimuli as we used to test human observers: ResNet-50 [15] as a representative for traditional ImageNet-trained CNNs and CLIP [7] as a representative for current SOTA models. For further details refer to Appendix A & B.



(a) Noise level = 0 (b) Noise level = 0.1 (c) Noise level = 0.2 (d) Noise level = 0.35

Figure 1: Systematic degradation of images with salt-and-pepper noise. We employed 4 different proportions of flipped pixels (0, 0.1, 0.2 or 0.35). For example, 0.2 means that 20% of the pixels are switched while 80% of the pixels remain untouched. Note that even though different degradation levels are shown for the same image, participants never encountered the same image more than once to avoid memory confounds.

3 Result I (psychophysical experiment): A four-year-old can outperform ResNet-50—object recognition robustness develops early

In order to assess the development of human OOD robustness we measured classification accuracy as a function of degradation level for all age groups and models. To compare the relative decrease in accuracy between the different observers and models we further normalised classification accuracy with respect to the initial accuracy at degradation level zero. The results are shown in Figure 2.

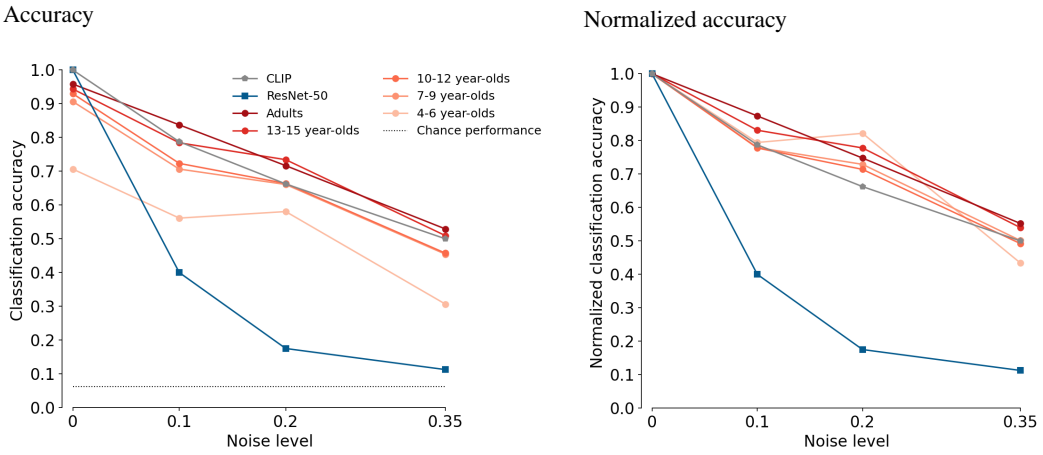


Figure 2: Classification accuracy (top 1) and normalised classification accuracy for different age groups, as well as for CLIP and ResNet-50. Both models were evaluated on the same images as human observers. The dotted line represents chance level performance of 6.25 % (100 % divided by the number of categories, which was 16). Normalised accuracy shows the change in accuracy relative to the initial accuracy at difficulty level zero of each age group or model, respectively.

Overall accuracy. Results indicate that object recognition of 4–6 year-olds is already remarkable robust against parametric image degradations. Overall, there is an increase of accuracy as a function of age. Between the 4–6 year-olds and the other age groups we find the largest increase in accuracy; 7–9 year-olds already show similar accuracy to adults. However, and most importantly, already 4–6 year-olds outperform ResNet-50 on very slightly distorted images, even in absolute performance. Object recognition robustness of young children thus seems to be much more similar to that of adults than that of ResNet-50. This becomes even clearer when we look at normalised accuracy. Here, we see that relative to their initial accuracy, 4–6 year-olds already demonstrate an adult-like robustness trajectory. In other words, while the younger children are not yet as good at recognising objects as the older age groups (as reflected by gap in absolute classification accuracy), their relative robustness seems already adult-like (as reflected by similar normalised accuracy across difficulty levels). CLIP shows—in contrast to ResNet-50—high OOD robustness and a human-like robustness trajectory w.r.t. increasing image degradations. Thus, Figure 2 illustrates that human OOD robustness develops very early and is already in place by the age of four.

Confusion matrices. Moving from the overall classification accuracy (coarse-grained) to classification accuracy and error patterns on the category-level (more fine-grained), we find a similar pattern of results. Figure 3 shows confusion matrices of children, adults and both models for moderately distorted images. While ResNet-50 is biased towards a small subset of categories and shows a tendency to categorise images regardless of the ground truth category (similar to [2, 16]), this is not the case for human observers. Just like adults, young children show high accuracies along the negative diagonal and do not favour a small subset of categories when making errors. Additionally, young children and adults find similar categories easy to recognise (e.g., airplane, bicycle, car, elephant, etc.). Those similarities further indicate that adult-like object recognition robustness emerges early in life. Again CLIP is more similar to human observers (irrespective of age) than to a standard trained CNN. The full set of confusion matrices (all age groups and all noise levels) can be found in Appendix C.

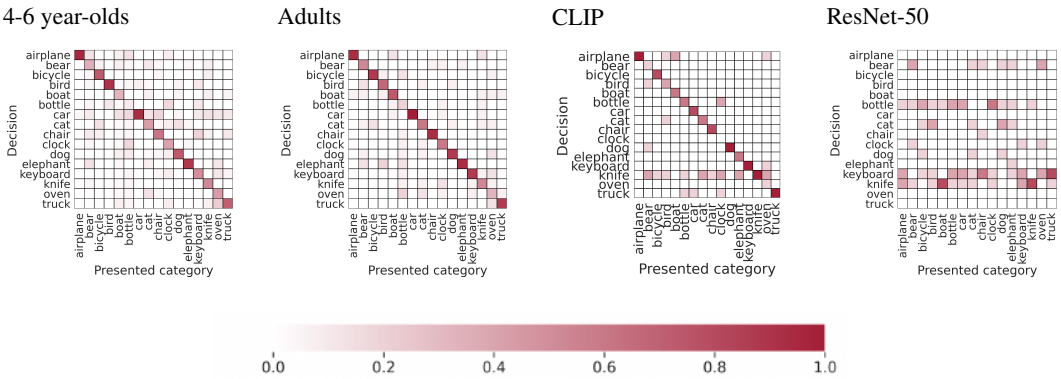


Figure 3: Confusion matrices for 4–6 year-olds, adults, CLIP and ResNet-50 in the salt-and-pepper noise condition at a signal-to-noise ratio of 4:1 (noise level = 0.2). Rows show classification decision of observers and CNNs and columns show the ground truth label of the presented category. Transparency of single squares within a matrix represent response probabilities (see color bar at the bottom). Entries along the negative diagonal represent correct responses; entries which are off the negative diagonal indicate errors.

4 Result II (back-of-the-envelope calculation): Human robustness doesn't require seeing billions of images during lifetime

So far, we have seen that object recognition emerges early and is largely in place by the age of four—but in order to put these results into context and relating them to models trained on large-scale datasets, we will need to estimate the number of "images" that young children are exposed to during their lifetime.

Basic idea. First, we performed a back-of-the-envelope calculation to attain the number of input images children are exposed to at different points during development. For each of those time points, we then looked at the corresponding human robustness and compared it to the OOD robustness of ResNet-50, trained on the same amount of images.

Estimating number of input images. During development, wake time gradually increases as a function of age; whereas on average, 0–1 year-olds are awake for 11.5 hours a day, adults are awake for 16.5 hours [17]. We took the mean age of each tested age group and calculated the total accumulated wake time for this particular age in seconds. To estimate input images, we then assumed one distinct input image every two seconds of wake time.[1] The number of input images for ResNet-50 was calculated by summing images over training epochs. During one epoch—5005 batches of 256 images—about 1.3M images are fed to the network. Standard training for CNNs consists of 90 epochs, which results in a total of 115M images. The classification robustness of ResNet-50 was evaluated every four epochs.

Estimating robustness. As a global measurement of classification robustness we calculated the mean classification accuracy over all moderately and heavily distorted images (noise level 0.2 and 0.35) for each age group (humans) or epoch (ResNet-50). Table 1 shows the robustness of children and ResNet-50 as a function of input images.

Interpretation. These back-of-the-envelope calculations show that, for a given number of input images, there are stark differences in OOD robustness between humans and a standard model trained on ImageNet (ResNet-50). This indicates that human OOD robustness may not require seeing billions

[1] Surely, the results depend heavily on this assumption. However, we are confident that our numbers are more of a conservative upper bound than a lower bound—simply imagine how much you would have to move around in order to get a completely different image every 2 seconds. Furthermore, human learning is most often only semi-supervised (i.e., the developing child is only sometimes provided with explicit labels; e.g, see [18]). In contrast, standard CNNs' input images are always labeled. Thus, in some sense one input image is "worth" more for a CNN than for a human.

of different images during lifetime. In other words, the human brain does not—in contrast to SOTA models—rely on a large-scale visual input to achieve high OOD robustness. Instead, the human brain may employ a smarter choice of information representation and processing optimized during evolution.

Table 1: Robustness of children and ResNet-50 as a function of input images. Each row contrasts object classification robustness of children and ResNet-50 given the same amount of input images. For example, at the age of four, a child was exposed to about 50M images, which corresponds to about 40 epochs of standard ImageNet training. Keeping the amount of cumulative input images constant like this, we can see a quite large difference in robustness between children and ResNet-50 (.44 vs .09). This indicates that human robustness develops early, and may not require seeing billions of images during lifetime. See Figure 4 in the Appendix for a full visualisation of the results.

Cumulative images	Age group	Epochs	Robustness children	Robustness ResNet-50
50M	4–6	40	**.44**	.09
77M	7–9	60	**.56**	.16
106M	10–12	84	**.56**	.14

5 Discussion

Summary. First we investigated the emergence of human OOD robustness during development. In order to do so, we employed a well-established psychophysical image-category identification task. Until now, investigations of the developmental trajectory of object recognition beyond two years of age have been scarce [11–13], and to our knowledge, this is the first study to directly compare children, adolescents, and different neural networks in a core object recognition task.[2] Second, we estimated the number of 'images' that young children have been exposed to during their lifetime. This showed that while large-scale training is very effective in increasing model OOD robustness, large-scale experience may not be required given that human robustness develops very early.

Limitations. The employed experimental set-up did not allow to test children younger than four years of age. The present study shows that human OOD robustness is largely in place by the age of four; it would be interesting to push this further and investigate the developmental trajectory during infancy to further disentangle the influence of evolution vs. lifetime experience. Especially regarding models without a convolutional architecture, it would be interesting to see at which point during training OOD robustness matches human robustness.

Conclusion and outlook. The results presented here suggest that the emergence of human OOD robustness takes place very early in development and does not seem to be dependent on large-scale visual input. That is, large-scale training might be sufficient, but not necessary for OOD robustness. It is more likely that superior information representation and processing—and, perhaps, multi-modal input from vision and touch, or active explorative behaviour of infants—is the crucial ingredient to human OOD robustness. Provided with a similar sized input dataset, even 4-year-olds easily outperform a standard trained CNNs on distorted images. Even though the current trend to train models with ever-larger datasets may close the longstanding robustness gap between humans and neural networks [6], it does not seem to capture the learning process by which human object recognition becomes robust. As a proof of concept, humans demonstrate that it is possible to learn much from little data—thus, much may be gained from smarter models, not just larger datasets.

References

[1] Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–7. IEEE, 2017.

[2] There is one unpublished investigation comparing children and CNNs that was presented at the 20th Annual Meeting of the Vision Sciences Society [19]. The findings presented there are in line with the results of the current study. However, the current study extends these findings by covering a larger age range, using naturalistic images with parameterized distortions and more response categories.

[2] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, 31:7538–7550, 2018b.

[3] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[4] Hongru Zhu, Peng Tang, Jeongho Park, Soojin Park, and Alan Yuille. Robustness of object recognition under extreme occlusion in humans and computational models. *arXiv preprint arXiv:1905.04598*, 2019.

[5] Jonas Kubilius, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Can deep neural networks rival human ability to generalize in core object recognition. *Cognitive Computational Neuroscience, 2018a. URL https://ccneuro. org/2018/Papers/ViewPapers. asp*, 2018.

[6] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *arXiv preprint arXiv:2106.07411*, 2021.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[10] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[11] Mayu Nishimura, Suzy Scherf, and Marlene Behrmann. Development of object recognition in humans. *F1000 Biology Reports*, 1, 2009.

[12] Linda B Smith. From fragments to geometric shape: Changes in visual object recognition between 18 and 24 months. *Current Directions in Psychological Science*, 18(5):290–294, 2009.

[13] Ingo Rentschler, Martin Jüttner, Erol Osman, Alexander Müller, and Terry Caelli. Development of configural 3d object recognition. *Behavioural Brain Research*, 149(1):107–111, 2004.

[14] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

[15] 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.

[16] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. On the surprising similarities between supervised and self-supervised models. *arXiv preprint arXiv:2010.08377*, 2020.

[17] B Thorleifsdottir, JK Björnsson, B Benediktsdottir, TH Gislason, and H Kristbjarnarson. Sleep and sleep habits from childhood to young adulthood over a 10-year period. *Journal of Psychosomatic Research*, 53(1):529–537, 2002.

[18] Alexander LaTourrette and Sandra R Waxman. A little labeling goes a long way: Semi-supervised learning in infancy. *Developmental Science*, 22(1):e12736, 2019. doi: 10.1111/desc.12736.

[19] Vladislav Ayzenberg and Stella Lourenco. Young children outperform feed-forward and recurrent neural networks on challenging object recognition tasks. *Journal of Vision*, 20(11): 310–310, 2020. doi: 10.1167/jov.20.11.310.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 5.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We don't foresee any negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] The reported experiments are approved by the ethics committee of our institution
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix A.3.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Appendix A.2
   (c) Did you include any new assets either in the supplemental material or as a URL? Due to the ImageNet permissions we are not allowed to share the employed stimuli publicly but we will be happy to provide them upon request.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The used data do not contain personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix A.3.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] There were no potential participant risks. The study was approved by the institutional ethical review board, participation was voluntary, and participants provided (parental) written consent.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix A.5

References

Appendix

We here report experimental details regarding the conducted psychophysical experiment (A), describe details regarding models, training and evaluation (B), and provide the full set of confusion matrices for all age groups and noise levels (C).

Figure 4 visualises average classification robustness as a function of cumulative number of input images for all age groups and all evaluated epochs.

A Experimental details regarding psychophysical experiments

A.1 Changes to the initial experimental paradigm

The methods used in this study are adapted from a series of psychophysical experiments conducted in the Wichmann-lab [2]. Here, we largely preserved this experimental procedure and relied on similar stimuli and setup. However, we adapted the following aspects to get a paradigm suitable to test children: We introduced a certain degree of gamification (see Section A.4), added more breaks, and suspended the fixed length of the experiment in order to increase data quality. Children who were motivated were allowed to continue the task, whereas children who were exhausted more quickly were given the possibility to terminate the task after each block of stimuli presentation (20 trials). Additionally, we slightly increased the stimulus presentation duration (from 200 ms to 300 ms) and only used stimuli which were correctly recognised by at least two adults in the previous studies. Finally, children were not tested in our visual psychophysics laboratory but in a quiet place in their school.

A.2 Dataset

As stimuli we used a subset of images (256×256 pixels) from the 16-class-ImageNet [2], which is a subset of ImageNet [20, license & download at `https://image-net.org/download.php`]. The 16-class-ImageNet database groups a subset of fine-grained ImageNet categories into 16 entry-level categories using the WordNet hierarchy [14]. In doing so it allows for easy comparison between humans and neural networks. A subset of 521 stimuli—stimuli that were correctly classified by at least two adults in prior experiments—served as a starting point for the present study. We chose this subset because we expected children's performance and motivation to be weaker compared to adults. Therefore, we felt that choosing a subset at least not containing exceedingly ambigious or even "impossible" images would prevent children from getting frustrated too quickly. We then randomly sampled 320 images (20 for each of the 16 categories) to be manipulated in the next step. A certain proportion of pixels were either set to a gray value of 1 or 0. This manipulation is often referred to as *salt and pepper noise*. We manipulated the images to four different degrees, resulting in 4 different difficulty levels, which corresponded to 4 different proportions of flipped pixels (0, 0.1, 0.2 or 0.35). For example, 0.2 means that 20% of the pixels are switched while 80% of the pixels remain untouched. The higher the proportion of switched pixels, the more difficult it is to recognise the content (see Figure 1 for an example).

A.3 Instructions and procedure

Each trial consisted of several phases. First, we presented an attention grabber inspired by an expiring clock (solid white circle which empties itself within 600 ms) in the center of the screen. We chose a moving stimulus instead of a more commonly used fixation cross in order to compensate for possible weaker attention in children. Second, the target image was shown in the center of the screen for 300 ms, followed by a full-contrast pink noise mask ($1/f$ spectral shape) of the same size and duration. Next, the screen turned blank and participants were required to indicate their answers. They did this by physically pointing to one of 16 icons corresponding with the 16 entry-level categories on a laminated DIN A4 sheet arranged in a 4×4 grid (icon size: 3×3 cm). We chose this physical response surface mainly for time efficacy (having 4-year-olds handle a computer mouse by themselves can be a lengthy undertaking). Next, the 16 icons appeared on screen, and the experimenter recorded the response using a wireless computer mouse. As in the experiments

conducted by Geirhos et al., our icons were a modified version of the ones from the MS COCO website (`https://cocodataset.org/#explore`).

All participants were tested in a quiet room in a single setting—either in their school (children, adolescents) or at home (adults). The experimental session started with some example images and 10 practice trials. For each category, we showed a prototypical example image in the center of the screen and asked participants to name the depicted object. The correct category was indicated by the subsequent presentation of the corresponding category icon. After completing all 16 examples, participants completed 10 practice trials that were identical to the experimental trials, except that undistorted color images were presented. If participants failed on two or more images, they had to do another round of 10 practice trials. Participants were instructed to respond as fast and accurately as possible, and to go with their best guess if unsure. Before the experimental trials started, a single distorted image (matched for the given condition) was shown and a short story-like explanation was given to justify why some of the subsequent images would be distorted ("Somebody spilled salt and pepper; that is why some of them look a bit strange."). Experimental trials were arranged in blocks containing 20 trials each. After each block participants received feedback and were asked whether they would like to continue, have a break or terminate the session. Adults were not explicitly asked if they want to terminate the session—but of course *all* participants were informed at the outset that they could abort the experiment at any given time. Participants could complete a maximum of 16 blocks (320 images). Code for the experiment can be found here.

Table 2: Descriptive statistics of participants and observations split by conditions. Sample size and quantity of observations (*n*), as well as mean (*M*) and standard deviation (*SD*) for age and trials within observer groups. Gender distribution (♂/♀) is given in percentages. Note that for adults, trial *M* equals the total number of trials in the respective condition and trial *SD* is zero because they had to go through all trials of that particular condition.

		Age within group				Trials	
Age group	*n*	*M*	*SD*	♂/♀	*n*	*M*	*SD*
4–6 year-olds	15	5.13	0.64	33/67	1240	62.66	54.96
7–9 year-olds	9	8.11	0.78	47/53	1840	204.44	112.60
10–12 years-olds	15	11	0.85	47/53	1880	125.33	71.9
13–15 years-olds	9	14.22	0.97	44/56	1280	142.22	82.12
Adults	3	28.33	5.51	33/67	960	320.00	0.00

A.4 Gamification

In order to increase motivation and make the experiment more appealing to children, we gamified several aspects of the experiment. In the beginning, participants could choose one of four characters (matched for gender) corresponding to four different roles: spy, detective, scientist or safari guide. All story lines had the same plot. The chosen character had to undergo a training session designed to improve his/her crucial skill. What the participants did not know was that the crucial skill, identifying objects as quickly and accurately as possible, was the same for all characters. After each trial, the chosen character was displayed at the foremost position of a progress bar indicating how far the participant had progressed in the current block (level). After each block, participants were provided with feedback designed to be perceptually similar to the display of a game score in an arcade game. There were three different types of scores. Participants received 10 `coins` as a reward for a finished block (not performance related). Additionally, for every two correctly recognised images, they received a `star` (performance-related). If they scored more than eight stars, they earned a special `emblem` matched for the chosen story character.[3] Figure 5 visualises different gamified elements.

A.5 Descriptive information and participant compensation

The total sample consisted of 46 children and three adults (60% female and 40% male). One child decided to cancel the study right after completing practice trials; therefore, no data was collected for

[3]The emblems were sunglasses for the spy, a magnifying glass for the detective, a microscope for the scientist, and a camera for the safari guide.

this participant. Descriptive information about the sample and observations is presented in Table 2. We recruited children from 17 different schools and the adult sample was recruited through private contacts. As an incentive for their participation, participants received a book of their choice. The total amount spent on participant compensation amounts to 433.3 (omitted for anonymous submission).

B Model, evaluation and training details

Results I. Both models were obtained through the `modelvshuman` Python-toolbox [6]. The ResNet-50 model was pre-trained on ImageNet and from the PyTorch model zoo version 1.4.0. The employed CLIP model has a ViT-B/32 backbone and is the most powerful CLIP model released till now. Evaluation was done employing the same toolbox with the following settings: Batch size = 16; number of workers = 4. Because images were grayscale, all three RGB channels were set to be equal to the grayscale image's single channel. Images were fed through the networks using a single feedforward pass.

Results II. We obtained a vanilla ResNet-50 model from the PyTorch model zoo version 1.4.0 and trained it from scratch on a SLURM cluster using a NVIDIA GeForce RTX 2080 Ti graphics card. The model was trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset which is a subset of ImageNet [20]. For the training we used the following hyperparameters: batch size = 256; number of workers = 20; epochs = 90; learning rate = 0.1; momentum = 0.9; weight decay = 1e-4. While training we saved all weights and biases every 4 epochs in order to evaluate robustness at different time points during training. Evaluation procedure was the same as for Results I.

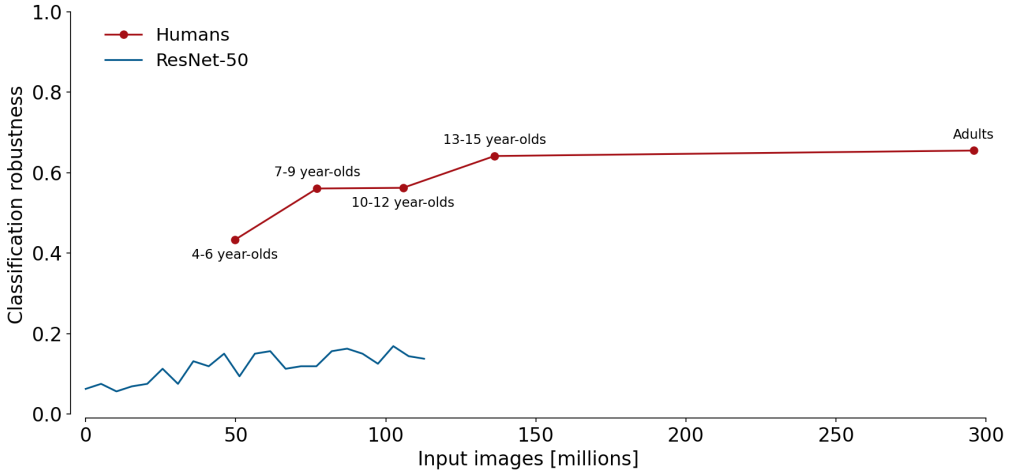
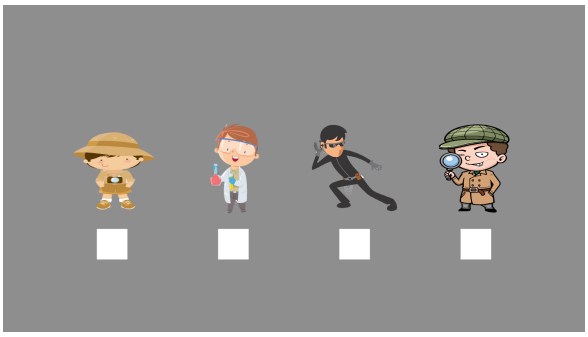


Figure 4: Average classification robustness as a function of cumulative number of input images. Classification robustness is calculated as the mean of classification accuracy on moderately and heavily distorted images. The number of input images for humans is estimated assuming one input image per two seconds of wake time and the number of input images for ResNet-50 is calculated by accumulating over training epochs (every 4 epochs).

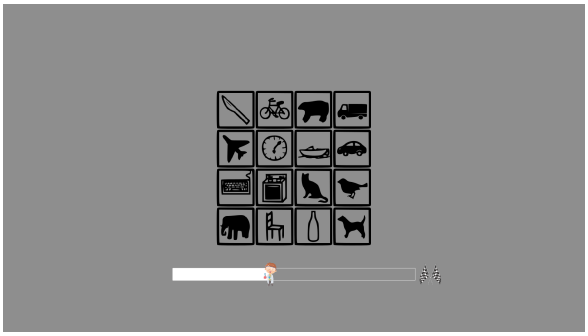
C Confusion matrices

Figure 6 shows confusion matrices for all observers across all noise levels. Up to and including noise level 0.2, all human observers show very similar confusion patterns. However, at noise level 0.35 the performance of 4–6 year-olds drops significantly. Nevertheless, response probabilities along the negative diagonal did not drop uniformly but remain high for some age-appropriate categories, such as such as `bear`, `bicycle`, `car`, or `elephant`. This indicates that young children´s weaker overall performance is not due to to a generally weaker ability to recognise objects but rather to weak performance on a subset of categories for which robust representations have not yet been acquired.

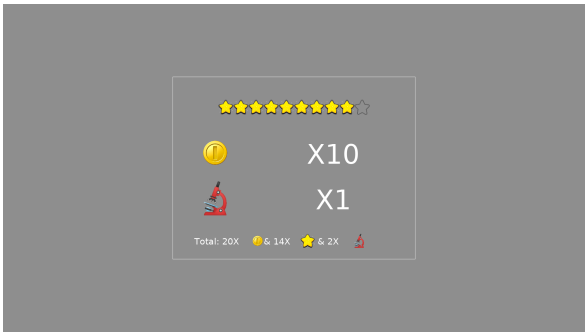
In comparison to human observers, ResNet-50 becomes biased towards a small subset of categories. That is, there is an emerging tendency to categorize images regardless of the ground truth category. At noise level 0.35 this biased behavior becomes even more pronounced. Almost all images are now categorised either as “bottle” or as “knife” by ResNet-50. CLIP on the other hand shows better overall performance—indicated by the higher response probabilities along the negative diagonal—as well as more human-like response patterns. There is even a similar subset of categories (e.g., `airplane`, `car`, or `keyboard`), which seems easy to recognise for both, CLIP and human observers. However, at noise level 0.35 CLIP starts to become biased towards a certain category (`knife`) as well. In other words, while CLIP shows—compared to ResNet-50—very human-specific behavior, it also seems to have a tendency which is typically associated with standard-trained CNNs. It would be interesting to check whether this tendency becomes more pronounced when the noise level is further increased.



(a) Character selection at the beginning of each session.

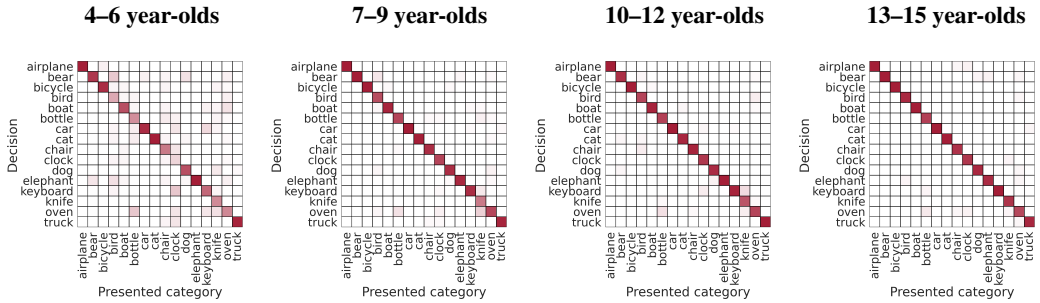


(b) Response screen with gamified progress-bar.

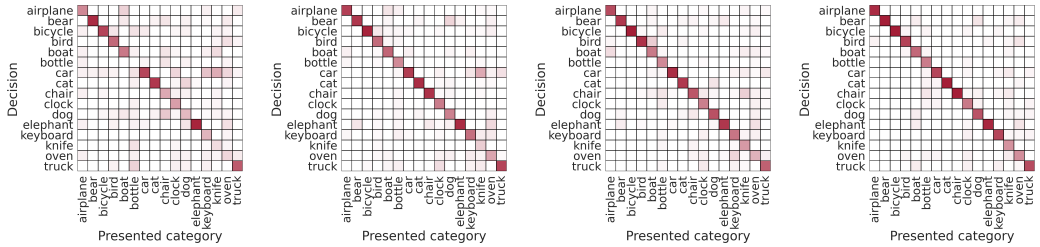


(c) Score display with star- and coin-scores and emblem.

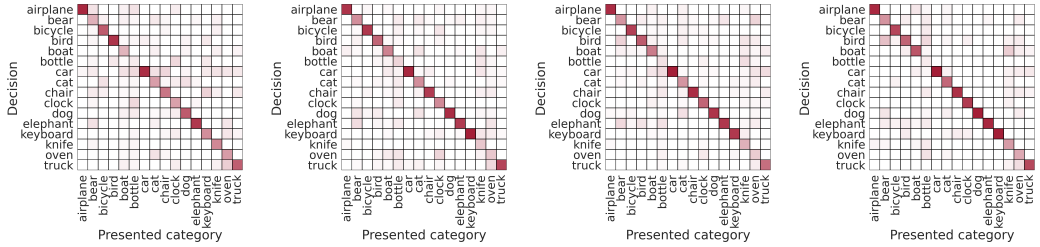
Figure 5: Screenshots at different time-points during the experiment.



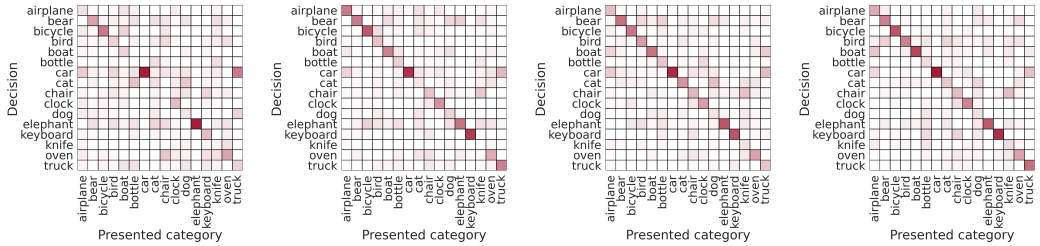
(a) Noise level = 0



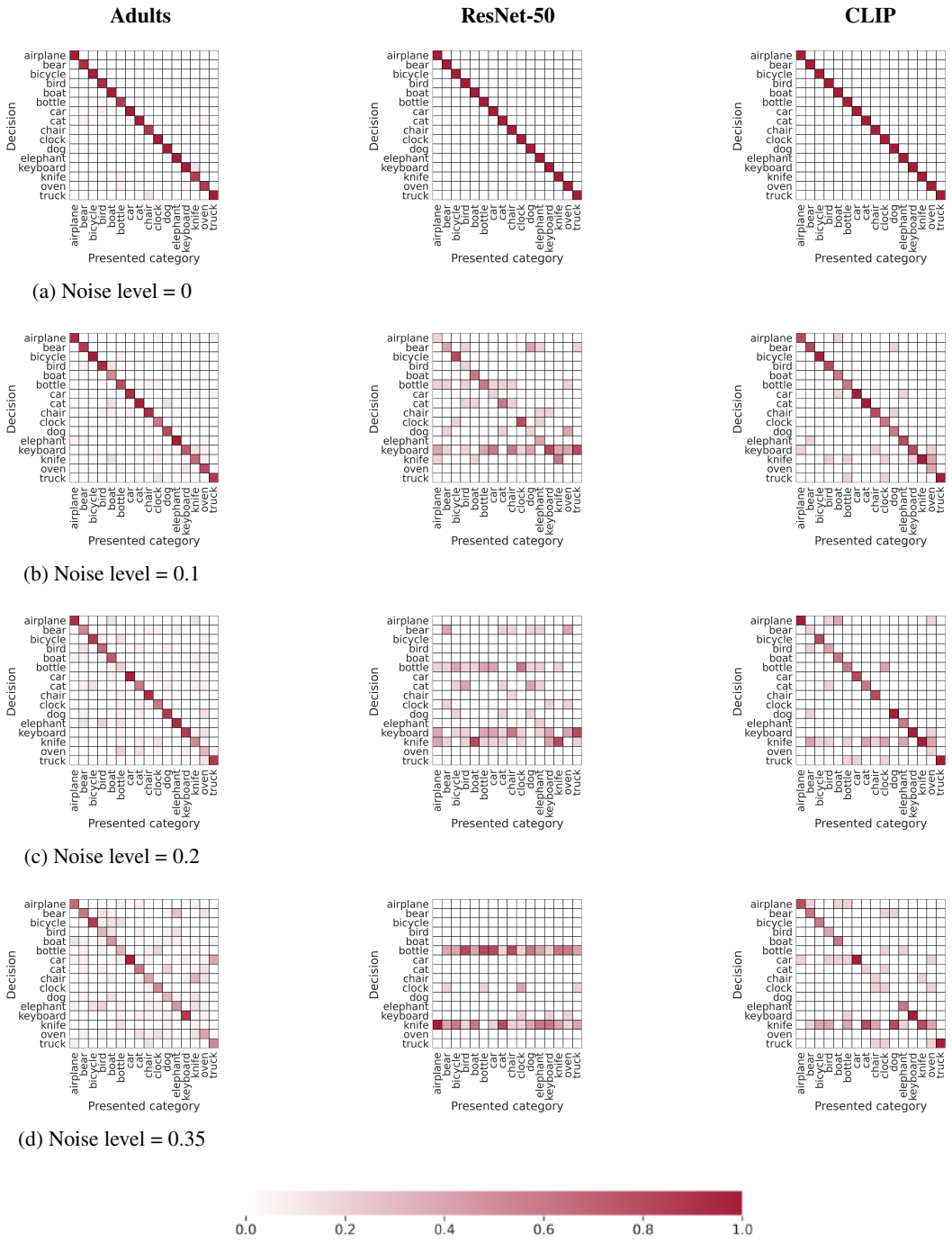
(b) Noise level = 0.1



(c) Noise level = 0.2



(d) Noise level = 0.35



(a) Noise level = 0

(b) Noise level = 0.1

(c) Noise level = 0.2

(d) Noise level = 0.35

Figure 6: Confusion matrices for all age groups and models across all noise levels (**a**-**d**). Rows show classification decision of observers and models and columns show the ground truth label of the presented category. Transparency of single squares within a matrix represent response probabilities (fully transparent = 0%, solid red = 100%). Entries along the negative diagonal represent correct responses; entries which are off the negative diagonal indicate errors.