Aligning human and machine vision

Abstract

The many successes of deep neural networks (DNNs) over the past decade have largely been driven by computational scale rather than insights from biological intelligence. Here, we explore if these trends have also carried concomitant improvements in explaining visual strategies underlying human object recognition. We do this by comparing two related but distinct properties of visual strategies in humans and DNNs: *where* they believe important visual features are in images and *how* they use those features to categorize objects. Across 85 different DNNs and three independent datasets measuring human visual strategies on ImageNet, we find a trade-off between DNN top-1 categorization accuracy and their alignment with humans. *State-of-the-art DNNs are progressively becoming less aligned with humans.* We rectify this growing issue by introducing the neural harmonizer: a general-purpose training routine that aligns DNN and human visual strategies while improving object classification performance. Our work represents the first systematic demonstration that the scaling laws [1–3] that are guiding DNN developments today have also produced worse models of human vision. We release our code and data to help the field build more human-like DNNs.

1 Introduction

In his bitter lesson [4], Rich Sutton wrote, "The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin." Deep learning has been the standard approach to object classification problems ever since the paradigm shifting success of AlexNet [5] on the ImageNet [6] benchmark a decade ago. As deep neural network (DNN) performance has continued to improve in the intervening years, Sutton's lesson has become more fitting than ever, with recent networks rivaling and likely outperforming humans on the benchmark [7] through brute-force computational scale: increasing the number of network parameters and number of images used for training orders-of-magnitude beyond AlexNet [1–3]. While the successes of so-called "scaling laws" are undeniable, this singular focus on performance in the field has side-stepped an equally important question that will govern the utility of object recognition models for brain sciences and industry applications alike: *are the visual strategies of DNNs aligned with human perception?*

The visual strategies that mediate object recognition in humans can be decomposed into two related but distinct processes: identifying *where* the important features for object recognition are in a scene, and determining *how* to integrate the selected features into a categorical decision [8, 9]. It has been known for nearly a century [10–13] that different humans attend to similar locations when asked to find and recognize objects. After selecting these important features, human observers are also consistent in how they use those features to classify objects. The inclusion of a few pixels in an image can be the difference between recognizing an object or not [9, 14].

Has the past decade of DNN development produced any models that are aligned with these human visual strategies for object recognition? Such a model could transform cognitive science by support-

ing a better mechanistic understanding of how vision works. More human-like models of object recognition would also resolve the problems with predictablity and interpretablity of DNNs [15–18], and control their alarming tendency to rely on "shortcuts" and dataset biases to perform well on tasks [19]. In this work, we perform the first large-scale and systematic comparison of the visual strategies of DNNs and humans for object recognition on ImageNet.

Contributions. In order to compare human and DNN visual strategies, we first turn to the human feature importance maps collected by Linsley et al. [20,21]. Their datasets, *ClickMe* and *Clicktionary*, contain maps of nearly 200,000 unique images in ImageNet that highlight the visual features humans believe are important for recognizing them. These datasets amount to a reverse inference on *where* important visual features are in ImageNet images (Fig. 1). We complement these datasets with new psychophysics experiments that directly test how important visual features are used for object recognition (Fig. 1). As DNN performance has increased on ImageNet, their alignment with human visual strategies captured in these datasets has worsened. This trade-off persists over 85 different DNNs representing all popular model classes – from those trained for adversarial robustness to those pushing the scaling laws in network capacity and training data. To summarize our findings:

- The trade-off between DNN object recognition accuracy and alignment with human visual strategies replicates across three unique datasets: *ClickMe* [20], *Clicktionary* [21], and our psychophysics experiments.

- We shift this trade-off with our neural harmonizer, a novel drop-in module for co-training any DNN to align with human visual strategies while also achieving high task accuracy. DNNs trained with the neural harmonizer learn visual strategies that are significantly more aligned with humans *than any other DNN we tested*.

- We release our data, code, and human psychophysics at `https://tinyurl.com/metapred` to help the field tackle the growing misalignment between DNNs and humans.

2   Related work

Do DNNs explain human visual perception?   Despite the continued success of DNNs on computer vision benchmarks, there are conflicting accounts on their ability to explain human vision. On the one hand, there's evidence that DNNs are becoming better models of humans on challenging visual tasks, such as recognizing objects obscured by noise [23]. On the other hand, there is a growing number of reports that DNNs struggle to explain perceptual phenomena in human vision like contextual illusions [24], perceptual grouping [19, 25, 26], and categorical prototypes [27]. Others

Figure 1: **Visual strategies of object recognition**. We investigate the alignment of human and DNN visual strategies in object classification. We decompose human visual strategies into descriptions of *where* important features are [20, 22], and *how* those features are integrated into visual decisions.

have found differences between human attention data and DNN models of attention [20, 28]. Moreover, similarities between DNNs and neural representations of objects have become less clear over the years. While the original theory was that model explanations of neural representations of object images improved alongside model classification accuracy [29], recent large-scale DNNs are worse at explaining neural data than older ones with lower ImageNet accuracy [30].

What are the visual strategies underlying human object recognition?   Ever since its inception, a goal of vision science has been to characterize the processes supporting human object recognition. It has been discovered that object recognition can be decomposed into different processing stages that emerge over time [8, 31–34], where the earliest stage is associated with processing through feedforward connections in the visual system, and the later stage is associated with processing through feedback connections. Since the DNNs used today rely on feedforward connections, it is likely that they are better models for that rapid feedforward phase of processing than the subsequent feedback

phase [33, 35]. Thus, to maximize the likelihood of alignment between DNN and human visual strategies, our experiments described in §3 test visual strategies underlying rapid feedforward object recognition in humans.

Most closely related to our work, and specifically analyses of *where* important visual features are in images, are studies of "top-down" image saliency. These involve asking participants to search for an object in an image, or specifically find the visual features needed to recognize an object [10–13, 20, 22, 36]. In our work, we complement these descriptions of *where* important features are in images with psychophysics testing *how* those features are used to classify objects.

Comparing visual strategies of humans and machines. As methods in explainable artificial intelligence have developed over the past decade, they have opened up opportunities for comparing the regions selected by humans and DNNs when solving tasks. Many of these comparisons have focused human image saliency measurements captured by eye tracking or mouse clicks during passive or active viewing [20, 22, 37–40]. Others have compared categorical representation distances [38, 41], or combined those distances with measures of human attention [28]. The most direct comparison between human and DNN visual strategies involved analyzing the minimal image patch needed to recognize an object [9, 42]. However, such studies were limited, comparing humans with several older DNNs using only a handful of images. To the best of our knowledge, the largest-scale comparison between human and DNN visual strategies involved the *ClickMe* dataset, which was used to compare visual regions preferred by humans for object classification to attention models [20]. What is noticeably missing from each of these studies is a large-scale study of how human and DNN alignment has changed as a function of performance.

Improving the correspondence between humans and machines. Inconsistencies between humans and DNN representations can be resolved by directly training models to act more like humans. DNNs have been trained to have more human-like attention, or human-like representational distances in their output layer [20, 38, 41, 43]. Here, we add to these successes with the neural harmonizer, a training routine designed to align the visual strategies (Fig. 1) between any two observers by minimizing the dissimilarity in their visual explanations.

3   Methods

Human feature importance datasets. We focus on the ImageNet dataset to compare the visual strategies of humans and DNNs for object recognition at scale. We rely on the two significant efforts for gathering feature importance data from humans on ImageNet: the *Clicktionary* [22] and *ClickMe* [20] games, which use slightly different methods to collect their data. Both games begin with the same basic setup: two observers play together to locate features in an object image that they believe are important for classifying it. As one of the observers selects important features, those features are filled into a blank canvas for the other observer to classify as quickly as possible. In *Clicktionary* [22], both players are humans, whereas in *ClickMe* [20], the player selecting features is a human and the player recognizing images is a CNN. In total, *Clicktionary* [22] contains feature importance maps for 200 images from the ImageNet validation set, whereas *ClickMe* [20] contains feature importance maps for 196,499 images from ImageNet training and validation sets. Thus, *ClickMe* has far more data than *Clicktionary*, but *Clicktionary* data has more reliable human feature importance data than *ClickMe*. By comparing human and DNN alignment on both of these datasets, we establish the robustness of our findings. Our experiments comparing human and DNN alignment compare feature importance maps from these datasets for the ImageNet validation set. As we describe in §4, we incorporate *ClickMe* feature importance maps for the ImageNet training set into our neural harmonizer routine to align human and DNN visual strategies.

Psychophysics participants and dataset. We complemented the feature importance maps from *Clicktionary* and *ClickMe* with psychophysics experiments on rapid visual categorization. We recruited 199 participants from Amazon Mechanical Turk (mturk.com) to complete the experiments. Participants viewed a psychophysics dataset consisted of the 100 animal and 100 non-animal images in the Clicktionary game taken from the ImageNet validation set [22]. We used the feature importance maps for each image as masks for the object images, allowing us to control the proportion of important features observers were shown when asked to recognize objects (Fig. 5a). We generated versions of each image that revealed anywhere between 1% to 100% (at log-scale spaced intervals) of the

Figure 2: **Human and DNNs rely on different features to recognize objects.** In contrast, our neural harmonizer aligns DNN feature importance with humans. We smooth the maps from both humans and DNNs with a Guassian kernel to improve visualization.

important object pixels (see SI §1 for details on mask generation). The total number of pixels revealed in each image was was equal for every image at a given level of image masking. The revealed image features were centered in each image to maximize the likelihood they fell on a partcipant's fovea. Each participant saw only one masked version of each object image.

Psychophysics experiment. Participants were instructed to categorize images in the psychophysics dataset as animals or non-animals as quickly and accurately as possible. Each experimental trial consisted of the following sequence of events overlaid onto a white background (SI Fig. 1): (*i*) a fixation cross displayed for a variable time (1,100–1,600ms); (*ii*) an image for 400ms; (*iii*) an additional 150ms of response time. In other words, the experiment forced participants to perform rapid object classification. They were given a total of 550ms to view an image and press a button to indicate its category (feedback was provided on trials in which responses were not provided within this time limit). Images were sized at 256 x 256 pixel resolution, which is equivalent to a stimulus size between approximately 5 – 11 degrees of visual angle across a likely range of possible display and seating setups participants used for the experiment. Similar paradigms and timing parameters have been shown to capture pre-attentive visual system processing [31, 44–46]. Participants provided informed consent electronically and were compensated $3.00 for their time ($\sim$ 10–15 min; $15.00/hr).

Models. We compared humans with 85 different DNNs representing the variety of approaches used in the field today: 58 CNNs trained on ImageNet [1, 47–59, 59–68], 4 CNNs trained on other datasets in addition to ImageNet (which we refer to as "CNN extra data") [1, 61, 69], 4 vision transformers [70–74], 6 CNNs trained with self-supervision [75, 76], and 13 models trained for robustness to noise or adversarial examples [77, 78]. We used pretrained weights for each of these models supplied by their authors, with a variety of licenses (detailed in SI §2), implemented in Tensorflow 2.0, Keras, or PyTorch.

4 Results

4.1 *Where* are diagnostic object features for humans and DNNs?

To systematically compare the visual strategies of object recognition for humans and DNNs on ImageNet, we first turned to the *ClickMe* dataset of feature importance maps [20]. In order to derive comparable feature importance maps for DNNs, we needed a method that could be efficiently and consistently applied to each of the 85 DNNs we tested with out any idiosyncratic hyperparameters. This led us to choose a classic method for explainable artificial intelligence, image feature saliency [79]. We prepared human feature importance maps from *ClickMe* by taking the average importance map across humans in every image that also appeared in ImageNet validation. We then recorded each model's top-1 accuracy on the entire ImageNet validation set and used Spearman's rank-correlation to measure the similarity between feature maps of humans and DNNs for each image. We also computed the maximum possible alignment for human feature importance maps by recording the correlation in every image between two randomly selected groups of participants. We stored the average correlation after repeating this procedure 1000 times, and then normalized each DNN's correlation with humans by this score.

There were dramatic qualitative differences between the features selected by humans and DNNs on ImageNet. In general, humans selected less contextual information and focused more on object details: the handle and opening of a bucket, the chassis of a snow mobile, or the frets of a banjo (Fig. 2). The DNN that was most aligned with humans, the DenseNet121, was still only 58% aligned with humans (Fig. 3).

Plotting the relationship between DNN ImageNet accuracy and alignment with humans revealed a striking trade-off: as model performance has improved beyond the DenseNet121, their alignment with humans has worsened (Fig. 3). For example, consider the ConvNext [1], which achieved the best top-1 accuracy in our experiments (85.8%), was only 22% aligned with humans – equivalent to the alignment of the BagNet33 [64] (63% top-1 accuracy). Overall, we observe that human and DNN alignment has considerably worsened since the introduction of these two models.

The neural harmonizer.   While scaling DNNs has had an immense impact on improving their performance on popular benchmark tasks, little has changed regarding the fundamental differences in architectures between DNNs and the human visual system [35]. While it's plausible that introducing more biological constraints into DNNs could improve their alignments with humans, there's evidence that doing so would hurt benchmark performance and require bespoke development for every different architecture [80–82]. *Is there a way to align any DNN's visual strategies with humans without hurting its performance?*

Such a general-purpose method for aligning human and DNN visual strategies should satisfy the following criteria: (***i***) Models trained with the method have feature importance maps that are aligned with humans across feature scales. While humans are scale-invariant [33, 83], DNNs typically are not; this constraint means that alignment must hold across scales. (***ii***) The method should work with any fully-differentiable network architecture. (***iii***) Task-accuracy of a model trained with the method should not be worse than a model trained without the method. We created the neural harmonizer to satisfy each of these criteria.

Let us consider a supervised classification problem with an input space $\mathcal{X}$ an output space $\mathcal{Y} \subseteq \mathbb{R}^c$ and a predictor function $\boldsymbol{f}_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ parameterized by θ, which maps an input vector $\boldsymbol{x} \in \mathcal{X}$ to an output $\boldsymbol{f}_\theta(\boldsymbol{x})$. We denote $\boldsymbol{g} : \mathcal{F} \times \mathcal{X} \rightarrow \mathcal{X}$ an explanation functional that, given a predictor $\boldsymbol{f}_\theta \in \mathcal{F}$ and an input, returns a feature importance map $\phi = \boldsymbol{g}(\boldsymbol{f}_\theta, \boldsymbol{x})$. Here, we focus on DNN saliency as our method for computing feature importance in DNNs, but the method can work with any differentiable network explanation method (see SI §2 for details).

To satisfy criterion (***i***), the neural harmonizer will enforce alignment across feature importance map scales. Let $\mathcal{P}_i(.)$ be the function mapping a feature importance map ϕ to it's representation in the N levels of a Gaussian pyramid, with $i \in \{1, ..., N\}$. The function $\mathcal{P}_i(\phi)$ is computed by downsampling $\mathcal{P}_{i-1}(\phi)$ using a Gaussian kernel, with $\mathcal{P}_1(\phi) = \phi$. We then seek to minimize $\sum_i^N ||\mathcal{P}_i(\boldsymbol{g}(\boldsymbol{f}_\theta, \boldsymbol{x})) - \mathcal{P}_i(\phi)||^2$, which will align feature importance maps between humans and DNNs at every scale of the pyramid. Furthermore, because this approach is fully differentiable, it also partially satisfies criterion (***ii***).

Figure 3: **The trade-off between DNN performance and alignment with human feature impor-tance from *ClickMe* [20]**. Human feature alignment is the mean Spearman correlation between human and DNN feature importance maps, normalized by the average inter-rater alignment of humans. The VGG16 (left) and the ResNet50 (right) trained with the neural harmonizer are more accurate and aligned than versions of those models trained only for classification. Aarows highlight the shift in performance per model.

To completely satisfy criterion (***ii***), the neural harmonizer should work well with training routines designed for large-scale computer vision challenges like ImageNet. This means the neural harmonizer must avoid optimization issues at scale. To do this, we need a way of comparing feature importance maps between humans and DNNs that is invariant to the norm of either map. We therefore standard-ize feature importance maps from humans and DNNs before comparing them, and only measure alignment on the most important areas of the image for each observer. Formally, let $z : \mathcal{X} \to \mathcal{X}$ be a standardization function over feature importance maps that takes the mean and standard deviation computed of each map ϕ such that $z(\phi)$ has 0 activation on average and unit standard deviation. To focus alignment on important regions, let $z(\phi)^+$ denote the positive part of the standardized explanation $z(\phi)$. Finally, to satisfy criterion (***iii***) and train models that are at least as accurate as those trained without the neural harmonizer, we include the familiar cross entropy objective to yield the complete neural harmonizer loss:

$$\mathcal{L} = \mathcal{L}_{CCE}(\boldsymbol{f}_\theta, \boldsymbol{x}, \boldsymbol{y}) + \lambda_1 \sum_i^N ||\boldsymbol{z}(\mathcal{P}_i((\boldsymbol{g}(\boldsymbol{f}_\theta, \boldsymbol{x}))^+ - \boldsymbol{z}(\mathcal{P}_i(\phi))^+||^2 + \lambda_2 \sum_i \theta_i^2 \qquad (1)$$

Training. We trained two different DNNs with the neural harmonizer, ResNet50 and VGG16, which were selected because they are popular networks that sit on the boundary of the trade-off between DNN performance and alignment with humans. Models were trained using the neural harmonizer to optimize classification performance on ImageNet and feature importance map alignment with human data from *ClickMe*. We trained models on all images in the ImageNet training set, but because *ClickMe* only contains human feature importance maps for a portion of those images, we computed the classification loss but not the alignment loss in Eq. (1) for images without importance maps. Models were trained using 8 cores V4 TPUs on the google cloud platform, and training lasted approximately one day. Models were trained with an augmented ResNet training recipe (built from https://github.com/tensorflow/tpu/). Models were optimized with SGD and momentum over batches of 512 images, a learning rate of 0.3, and label smoothing [84]. Images were augmented with random left-right flips, mixup [85]. The learning rate was adjusted over the course of training with a schedule that began with an initial warm-up period of 5 epochs and then decaying according

to a cosine function over 90 epochs, with decay at step 30, 50 and 80. The number of steps per epoch was multiplied by a factor of 3. We validated that a ResNet50 and VGG16 trained with these hyperparameters and schedule using standard cross-entropy (instead of the neural harmonizer) matched published performance.

The neural harmonizer aligns human and DNN visual strategies. Returning to the observed trade-off between ImageNet accuracy and model alignment with *ClickMe* human feature importance maps, we found that training models with the neural harmonizer broke this trade-off. All three of our neural harmonizer models were significantly more aligned with feature importance maps and performed better on ImageNet (Fig. 3). The changes in feature importance are dramatic: a ResNet50 trained with the neural harmonizer had feature importance maps that are far less reliant on context (Fig. 1) and 20% more aligned with humans (Fig. 3; 69.17% alignment) than the standard ResNet50. The same model also performed 2% better in top-1 accuracy, matching the performance of a DenseNet 121, without any changes to its architecture. A VGG16 trained with the neural harmonizer had feature importance maps that were 269% more aligned with humans (94.45% alignment) and performed 3% better in top-1 accuracy than a standard VGG16.

Figure 4: **The trade-off between DNN performance and alignment with human feature importance from *Clicktionary* [22].** Human feature alignment is the mean Spearman correlation between human and DNN feature importance maps, normalized by the average inter-rater alignment of humans. The VGG16 (left) and the ResNet50 (right) trained with the neural harmonizer are more accurate and aligned than versions of those models trained only for classification. Aarows highlight the shift in performance per model.

Clicktionary. To test if the trade-off between DNN ImageNet accuracy and alignment with humans is general phenomenon we next turned to *Clicktionary* [22]. Indeed, after comparing DNNs and humans using the same approach we used for *ClickMe*, we observed a similar trade-off: as DNN accuracy has improved on ImageNet, feature importance alignment with participants in *Clicktionary* has worsened (Fig. 4). As with *ClickMe*, DNNs trained with the neural harmonizer shift the accuacy-alignment trade-off on this dataset.

4.2 *How* do humans and DNNs integrate diagnostic object features into decisions?

The trade-off we discovered between DNN accuracy on ImageNet and alignment with human visual feature importance suggests that the two use different visual strategies. However, there is potential for an even deeper problem. Even if two observers deem the same regions of an image as important for recognizing it, there is no guarantee that they use the selected features in the same way to render

Figure 5: **Comparing *how* humans and DNNs use visual features during object recognition**. **(a)** Humans and DNNs categorized ImageNet validation images as animals or non-animals. The images revealed only a portion of the most important visual features according to the *Clicktionary* game [86]. **(b)** There was a trade-off between DNN top-1 accuracy on ImageNet and alignment with human visual decision making. **(c)** A state-of-the-art DNN like the ViT learned a different strategy for integrating visual features into decisions than humans or a model trained with the neural harmonizer (in this case, a ResNet50).

their decisions. We posit that if two observers have aligned visual strategies, the will agree on both *where* important features are in an image and *how* they use those features for decisions.

We developed a psychophysics experiment to measure how different humans use features in ImageNet images to recognize objects. Participants viewed versions of these images where only a proportion of the features that were deemed most important in the *Clicktionary* game were visible (Fig. 5a), and

had to accurately detect whether or not the image contained an animal within 550ms. This rapid categorization paradigm controlled for confounds in our comparisons with the feedfoward DNNs we tested by minimizing the likelihood that participants relied on anything but feedforward processing. Participant performance on this experiment revealed how their feedforward processes integrated features into accurate decisions. We performed the same experiment on DNNs as we did on humans, recording animal vs. non-animal decisions according to whether or not the most probable category in the model's 1000-category output was an animal. Because the experiment was speeded, human participants did not achieve perfect accuracy. Thus, we normalize performance for humans and DNNs in order to compare the rate at which each integrates featuers into visual decisions.

We discovered a similar trade-off between ImageNet accuracy and alignment with human visual decision making in this experiment as we did in *ClickMe* and *Clicktionary* (Fig. 5b). Indeed, the model that was most aligned with human decision-making – the BagNet33 [64] – only achieved 63.0% accuracy on ImageNet. Surprisingly, models trained with the neural harmonizer broke this trend, particularly the harmonized ResNet50 (Fig. 5b, top-right), despite no explicit constraints in that procedure which forced consistent decision-making with humans. In contrast, a performant DNN like ViT (85.8% accuracy) integrates information more slowly than humans or models trained with the neural harmonizer (Fig. 5c).

5 Conclusion

Models that reliably categorize objects like humans do would shift the paradigms of the cognitive sciences and artificial intelligence. But despite continuous progress over the past decade on the ImageNet benchmark, we find that DNNs are becoming *worse* models of human vision. Our solution to this problem, the neural harmonizer routine, can be applied to any DNN to align their visual strategies with humans and, at least in the models we tested, even improve performance.

The scaling laws that are governing DNN development today are choosing to forgo human alignment in exchange for incredible performances on nearly every benchmark in artificial intelligence. We encourage the field to re-analyze the costs and benefits of this exchange, particularly in light of the growing worries about DNNs leveraging shortcuts and dataset biases to achieve high performance [19]. Alignment with human vision need not be exchanged with performance if DNNs are trained with our neural harmonizer. Our project code (`https://tinyurl.com/metapred`) can be used to incorporate the neural harmonizer in any new DNN created and measure its alignment with humans on the datasets we describe in this paper.

Limitations. One possible explanation for the misalignment between DNNs and humans that we observe is that recent DNNs have achieved superhuman accuracy on ImageNet. Superhuman DNNs have been described in biomedical applications [87, 88] where there's definitive biological ground-truth labels, but ImageNet labels are noisy, making it unclear if such an achievement is laudable (`https://labelerrors.com/`). Thus, an equally likely explanation is that the continued improvements of DNNs at least partially reflect their exploitation of shortcuts in ImageNet [19]. We leave experiments that can adjudicate between these possibilities to future work, and recommend practitioners use the neural harmonizer to ensure trustworthy behavior in DNNs.

Broader impacts. A persistent issue in the field of artificial intelligence is the tendency of models to exploit dataset biases. With a growing trend towards automating every facet of society with DNNs, models exploiting such biases have the potential to learn strategies that systematically oppress groups of people. A central theme of our work is that there's elements of human vision that are not captured by DNNs, particularly those which follow the scaling laws which have been so embraced by industry leaders. Forcing DNNs to rely on similar visual strategies as humans could represent a scalable path forward to correcting the insidious biases which have assailed applications of artificial intelligence. At the very least, this strategy would make DNNs more predictable and interpretable for their human creators, to make it easier to filter out bad models.

References

[1] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. (January 2022)

[2] Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. (June 2021)

[3] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. (January 2020)

[4] Sutton, R.: The bitter lesson. (May 2019)

[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems. Volume 25., Curran Associates, Inc. (2012)

[6] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. (June 2009) 248–255

[7] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. (September 2014)

[8] DiCarlo, J.J., Zoccolan, D., Rust, N.C.: How does the brain solve visual object recognition? Neuron **73**(3) (February 2012) 415–434

[9] Ullman, S., Assif, L., Fetaya, E., Harari, D.: Atoms of recognition in human and computer vision. Proc. Natl. Acad. Sci. U. S. A. **113**(10) (March 2016) 2744–2749

[10] Buswell, G.T.: How people look at pictures: a study of the psychology and perception in art. **198** (1935)

[11] Yarbus, A.L.: Eye Movements and Vision. Springer US

[12] Posner, M.I.: Orienting of attention. Q. J. Exp. Psychol. **32**(1) (February 1980) 3–25

[13] Mannan, S.K., Kennard, C., Husain, M.: The role of visual salience in directing eye movements in visual object agnosia. Curr. Biol. **19**(6) (March 2009) R247–8

[14] Gruber, L.Z., Ullman, S., Ahissar, E.: Oculo-retinal dynamics can explain the perception of minimal recognizable configurations. Proc. Natl. Acad. Sci. U. S. A. **118**(34) (August 2021)

[15] Fel, T., Colin, J., Cadene, R., Serre, T.: What I cannot predict, I do not understand: A Human-Centered evaluation framework for explainability methods. (December 2021)

[16] Fel, T., Cadene, R., Chalvidal, M., Cord, M., Vigouroux, D., Serre, T.: Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. (November 2021)

[17] Fel, T., Ducoffe, M., Vigouroux, D., Cadene, R., Capelle, M., Nicodeme, C., Serre, T.: Don't lie to me! robust and efficient explainability with verified perturbation analysis. (February 2022)

[18] Fel, T., Vigouroux, D., Cadène, R., Serre, T.: How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. (September 2020)

[19] Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11) (November 2020) 665–673

[20] Linsley, D., Shiebler, D., Eberhardt, S., Serre, T.: Learning what and where to attend with humans in the loop. In: International Conference on Learning Representations. (2019)

[21] Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017) 936–944

[22] Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., Serre, T.: What are the visual features underlying human versus machine vision? In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). (October 2017) 2706–2714

[23] Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F.A., Brendel, W.: Partial success in closing the gap between human and machine vision. (June 2021)

[24] Linsley, D., Kim, J., Ashok, A., Serre, T.: Recurrent neural circuits for contour detection. International Conference on Learning Representations (2020)

[25] Kim*, J., Linsley*, D., Thakkar, K., Serre, T.: Disentangling neural mechanisms for perceptual grouping. International Conference on Representation Learning (2020)

[26] Linsley, D., Malik, G., Kim, J., Govindarajan, L.N., Mingolla, E., Serre, T.: Tracking without re-recognition in humans and machines. (May 2021)

[27] Golan, T., Raju, P.C., Kriegeskorte, N.: Controversial stimuli: Pitting neural networks against each other as models of human cognition. Proc. Natl. Acad. Sci. U. S. A. **117**(47) (November 2020) 29330–29337

[28] Langlois, T., Zhao, H., Grant, E., Dasgupta, I., Griffiths, T., Jacoby, N.: Passive attention in artificial neural networks predicts human visual selectivity. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W., eds.: Advances in Neural Information Processing Systems. Volume 34., Curran Associates, Inc. (2021) 27094–27106

[29] Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U. S. A. **111**(23) (June 2014) 8619–8624

[30] Schrimpf, M., Kubilius, J., Lee, M.J., Ratan Murty, N.A., Ajemian, R., DiCarlo, J.J.: Integrative benchmarking to advance neurally mechanistic models of human intelligence. Neuron **108**(3) (November 2020) 413–423

[31] Fabre-Thorpe, M.: The characteristics and limits of rapid visual categorization. Front. Psychol. **2** (October 2011) 243

[32] Roelfsema, P.R., Lamme, V.A., Spekreijse, H.: The implementation of visual routines. Vision Res. **40**(10-12) (2000) 1385–1411

[33] Serre, T., Oliva, A., Poggio, T.: A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. U. S. A. **104**(15) (April 2007) 6424–6429

[34] Kietzmann, T.C., Spoerer, C.J., Sörensen, L.K.A., Cichy, R.M., Hauk, O., Kriegeskorte, N.: Recurrence is required to capture the representational dynamics of the human visual system. Proc. Natl. Acad. Sci. U. S. A. **116**(43) (October 2019) 21854–21863

[35] Serre, T.: Deep learning: The good, the bad, and the ugly. Annu Rev Vis Sci **5** (September 2019) 399–426

[36] Koehler, K., Guo, F., Zhang, S., Eckstein, M.P.: What do saliency models predict? J. Vis. **14**(3) (March 2014) 14

[37] Jiang, M., Huang, S., Duan, J., Zhao, Q.: SALICON: Saliency in context. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015) 1072–1080

[38] Peterson, J.C., Abbott, J.T., Griffiths, T.L.: Evaluating (and improving) the correspondence between deep neural networks and human representations. Cogn. Sci. **42**(8) (November 2018) 2648–2669

[39] Lai, Q., Khan, S., Nie, Y., Shen, J., Sun, H., Shao, L.: Understanding more about human and machine attention in deep neural networks. (June 2019)

[40] Ebrahimpour, M.K., Falandays, J.B., Spevack, S., Noelle, D.C.: Do humans look where deep convolutional neural networks "attend"? In: Advances in Visual Computing, Springer International Publishing (2019) 53–65

[41] Roads, B.D., Love, B.C.: Enriching ImageNet with human similarity judgments and psychological embeddings. (November 2020)

[42] Funke, J., Tschopp, F.D., Grisaitis, W., Sheridan, A., Singh, C., Saalfeld, S., Turaga, S.C.: Large scale image segmentation with structured loss based deep learning for connectome reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. (2018) 1–1

[43] Boyd, A., Tinsley, P., Bowyer, K., Czajka, A.: CYBORG: Blending human saliency into the loss improves deep learning. (December 2021)

[44] Eberhardt, S., Cader, J.G., Serre, T.: How deep is the feature analysis underlying rapid visual categorization? In Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., eds.: Advances in Neural Information Processing Systems 29. Curran Associates, Inc. (2016) 1100–1108

[45] Kirchner, H., Thorpe, S.J.: Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. Vision Res. **46**(11) (May 2006) 1762–1776

[46] Muriel, B., Simon, T., Holle, K.: Rapid object categorization without conscious recognition: aneuropsychological study. J. Vis. **7**(9) (June 2007) 1033–1033

[47] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. (March 2021)

[48] Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. (May 2019)

[49] Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. (March 2020)

[50] Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for MobileNetV3. (May 2019)

[51] Simonyan, K., Zisserman, A.: Very deep convolutional networks for Large-Scale image recognition. (September 2014)

[52] Huang, L., Zhao, X., Huang, K.: GOT-10k: A large High-Diversity benchmark for generic object tracking in the wild. (October 2018)

[53] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. (December 2015)

[54] Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M., Smola, A.: ResNeSt: Split-Attention networks. (April 2020)

[55] Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2Net: A new Multi-Scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2) (February 2021) 652–662

[56] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (BiT): General visual representation learning. (December 2019)

[57] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted residuals and linear bottlenecks. (January 2018)

[58] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, Inception-ResNet and the impact of residual connections on learning. (February 2016)

[59] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. (December 2015)

[60] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. (October 2016)

[61] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. (February 2021)

[62] Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A., Le, Q.V.: Adversarial examples improve image recognition. (November 2019)

[63] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. (November 2016)

[64] Brendel, W., Bethge, M.: Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. (March 2019)

[65] Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C.: XNect: real-time multi-person 3D motion capture with a single RGB camera. ACM Trans. Graph. **39**(4) (July 2020) 82:1–82:17

[66] Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J.: Dual path networks. (July 2017)

[67] Wang, C.Y., Liao, H.Y.M., Yeh, I.H., Wu, Y.H., Chen, P.Y., Hsieh, J.W.: CSPNet: A new backbone that can enhance learning capability of CNN. (November 2019)

[68] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., Le, Q.V.: MnasNet: Platform-Aware neural architecture search for mobile. (July 2018)

[69] Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves ImageNet classification. (November 2019)

[70] d'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G., Sagun, L.: ConViT: Improving vision transformers with soft convolutional inductive biases. (March 2021)

[71] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. (December 2020)

[72] Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A.: MLP-Mixer: An all-MLP architecture for vision. (May 2021)

[73] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. (October 2020)

[74] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your ViT? data, augmentation, and regularization in vision transformers. (June 2021)

[75] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. (February 2020)

[76] Zeki Yalniz, I., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. (May 2019)

[77] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. (November 2018)

[78] Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., Madry, A.: Do adversarially robust ImageNet models transfer better? (July 2020)

[79] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. (December 2013)

[80] Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., Kreiman, G.: Recurrent computations for visual pattern completion. Proc. Natl. Acad. Sci. U. S. A. **115**(35) (August 2018) 8835–8840

[81] Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D.L., DiCarlo, J.J.: Brain-Like object recognition with High-Performing shallow recurrent ANNs. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., Garnett, R., eds.: Advances in Neural Information Processing Systems 32. Curran Associates, Inc. (2019) 12805–12816

[82] Schrimpf, M., Kubilius, J., Lee, M.J., Ratan Murty, N.A., Ajemian, R., DiCarlo, J.J.: Integrative benchmarking to advance neurally mechanistic models of human intelligence. Neuron **108**(3) (November 2020) 413–423

[83] Han, Y., Roig, G., Geiger, G., Poggio, T.: Scale and translation-invariance for novel objects in human vision. Sci. Rep. **10**(1) (January 2020) 1411

[84] Müller, R., Kornblith, S., Hinton, G.: When does label smoothing help? (June 2019)

[85] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. (October 2017)

[86] Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., Serre, T.: What are the visual features underlying human versus machine vision? (January 2017)

[87] Linsley, J.W., Linsley, D.A., Lamstein, J., Ryan, G., Shah, K., Castello, N.A., Oza, V., Kalra, J., Wang, S., Tokuno, Z., Javaherian, A., Serre, T., Finkbeiner, S.: Superhuman cell death detection with biomarker-optimized neural networks. Sci Adv **7**(50) (December 2021) eabf8142

[88] Lee, K., Zung, J., Li, P., Jain, V., Sebastian Seung, H.: Superhuman accuracy on the SNEMI3D connectomics challenge. (May 2017)