Improved accuracy in deep vision models does not come with better predictions of perceptual similarity

Fritz Günther (fritz.guenther@hu-berlin.de)
Department of Psychology, Humboldt-Universität zu Berlin, Unter den Linden 6
10099 Berlin, Germany

Marco Marelli (marco.marelli@unimib.it)
Department of Psychology, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1
20126 Milan, Italy

Marco Alessandro Petilli (marco.petilli@unimib.it)
Department of Psychology, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo 1
20126 Milan, Italy

Abstract

Over the last years, advancements in deep learning models for computer vision have led to a dramatic improvement in their image classification accuracy. However, models with a higher accuracy in the task they were trained on do not necessarily develop better image representations that allow them to also perform better in other tasks they were not trained on. In order to investigate the representation learning capabilities of prominent high-performing computer vision models, we investigated how well they capture various indices of perceptual similarity from large-scale behavioral datasets. We find that higher image classification accuracy rates are not associated with a better performance on these datasets, and in fact we observe no improvement in performance since GoogLeNet (released 2015) and VGG-M (released 2014). We speculate that more accurate classification may result from hyper-engineering towards very fine-grained distinctions between highly similar classes, which does not incentivize the models to capture overall perceptual similarities.

Keywords: vision models; computer vision; representation learning; visual similarity

Over the last decade, following the seminal work by Krizhevsky, Sutskever, and Hinton (Krizhevsky, Sutskever, & Hinton, 2012), computer vision models based on deep neural network architectures have become increasingly powerful, and nowadays achieve very high levels of performance (Byerly, Kalganova, & Ott, 2022; Szegedy, Ioffe, Vanhoucke, & Alemi, 2017). This performance is typically assessed on the very task used in model training, most often as the accuracy in image classification (using measures such as top-1 error or top-5 error; Russakovsky et al., 2015).

As these models achieve higher and higher performance in such scenarios, they also tend to become increasingly sophisticated and complex in terms of model architecture and the numbers of parameters to be estimated. However, this additional complexity does not necessarily imply that these models *generally* perform better, also on domains they are *not* trained on: such an approach runs the risk of having systems that are over-optimized for a particular (set of) tasks, without gaining much in terms of transfer and generalizability (Goodfellow, Bengio, & Courville, 2016).

These aspects play an important role in machine learning, often discussed under the label of *representation learning* (Goodfellow et al., 2016). However, the point is even more relevant when these systems are used as general-level vision models for research purposes. In that respect, an emerging line of research in the domains of computational neuroscience and cognitive science has started to investigate and employ computer vision models (originally designed and trained for image classification) as models for human visual representation and processing, with very promising results from recent studies (Battleday, Peterson, & Griffiths, 2021; Günther, Marelli, Tureski, & Petilli, 2023). These works also provide us with rich, large-scale datasets of human behavioral data that allow us to investigate to which extent current computer-vision models can serve as general-level vision models, with much wider scientific applications than being pure image classifiers (Cichy & Kaiser, 2019; Kriegeskorte, 2015; Lindsay, 2021). Following these developments, in the present study, we will systematically examine which models perform best when tested against a battery of behavioral datasets, and if such models also turn out to be the most complex and best-performing image classifiers.

Related Work

In other computational modelling domains such as the development of language models, human behavioral data have long been established as a gold standard for model evaluation (e.g., Baroni, Dinu, & Kruszewski, 2014). The most prominent example are ratings of word similarity, with widely-used datasets such as WordSim353 (Finkelstein et al., 2001), SimLex999 (Hill, Reichart, & Korhonen, 2015), or MEN (Bruni, Tran, & Baroni, 2014).

Analogously, ratings of image similarity are widely employed to evaluate and compare the performance of computer vision models. This includes pairs of different naturalistic images (Hebart, Zheng, Pereira, & Baker, 2020; Jozwik, Kriegeskorte, Storrs, & Mur, 2017; Peterson, Abbott, & Griffiths, 2018), as well as comparisons between real images and their distorted versions (R. Zhang, Isola, Efros, Shechtman, & Wang, 2018). In a recent study, Roads and Love (2021) collected similarity ratings for a very large collection of 50,000 ImageNet images, which were not only used for evaluation but also to enrich computer vision with participant-sourced information.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

More recently, Günther et al. (2023) released a collection of large-scale data sets, comprising rating data as well as online processing data in the form of response times, which were used to evaluate a VGG-based vision model (Chatfield, Simonyan, Vedaldi, & Zisserman, 2014). These will constitute the gold standard datasets for our present study, where we systematically evaluate the performance of a wide range of models against data that are cognitively relevant, but relatively atypical for the computer vision domain, and far from the tasks on which systems are typically optimized.

Datasets

We considered the following metrics from the datasets provided by Günther et al. (2023) (see this paper for details on data collection):

- **Ratings** of
  - **image similarity [IMG]** for 3,000 pairs of naturalistic ImageNet images (a total of 6,000 images from 2,228 different categories). Data were collected from 480 participants, with 30 observations per image pair.
  - ***visual* word similarity [WORD]** for 3,000 word pairs (image labels of the aforementioned 3,000 image pairs from 2,228 categories), where participants were asked to judge how similar of the objects denoted by the words (i.e., the word referents) *look like*. Thus, unlike other word-based ratings (Bruni et al., 2014; Finkelstein et al., 2001; Hill et al., 2015), these data focus on the visual domain. Data were collected from 480 participants, with 30 observations per word pair.
  - **typicality ratings [TYP]** for 7,500 word-image pairs (sets of 1,500 different image labels/categories and five images tagged with that label), where participants were asked to indicate the most and least typical image for the category denoted by the presented label. Data were collected from 902 participants, with 30 observations per word-image pair.

  All ratings were collected using the best-worst method (Hollis, 2018), so participants were always presented with a set of stimuli and asked to pick the most and least relevant for the given task. Responses were then scored on a continuous scale using the Value learning algorithm (Hollis, 2018). The datasets thus contain exactly one rating score between 0 (completely dissimilar) and 1 (identical) for each word pair in the WORD dataset and each image pair in the IMG dataset, and one score between 0 (very atypical) and 1 (very typical) for each word-image pair in the TYP dataset. Examples for items with very high and very low ratings are presented in Figure 1

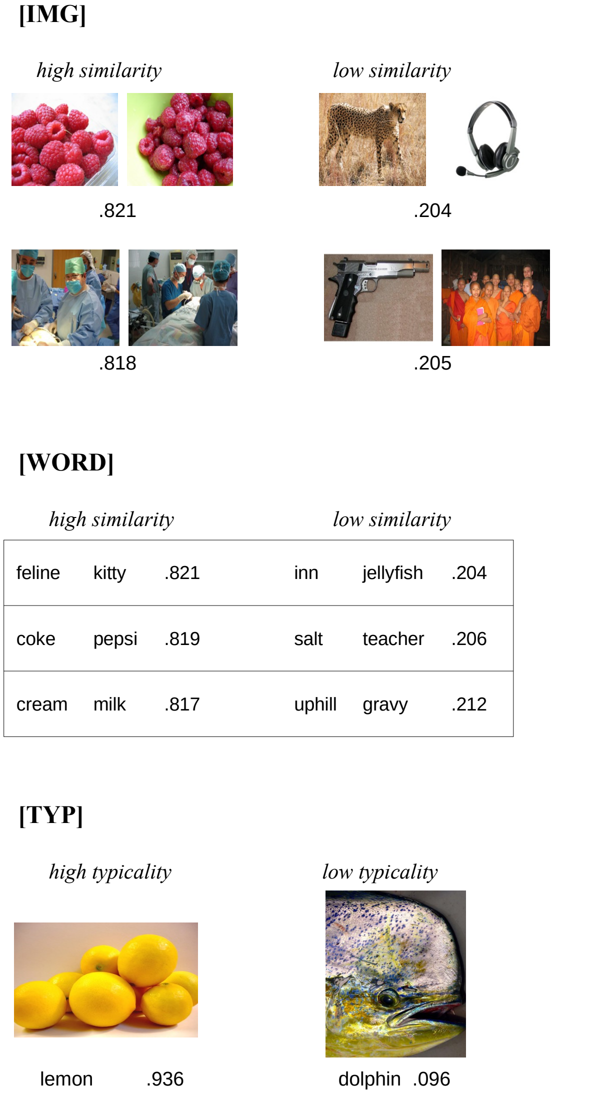

Figure 1: Examples for items with very high and very low rating values in the individual rating tasks. *Upper panel:* Image similarity ratings [IMG]; *middle panel:* word similarity ratings for visual similarity between the words denoted by the objects [WORD], *lower panel:* typicality ratings.

- **Processing time data**
  - **discrimination task [DIS]** for the same 3,000 image pairs of the IMG dataset. In a discrimination task, two stimuli (here: images) are presented in very rapid succession, and participants have to indicate whether they are identical or different by pressing one of two buttons (see Figure 2, upper panel for a schematic representation of an experimental trial). Responses are typically *slower* for more visually similar stimuli, which are harder to discriminate from the actual stimulus. Data were collected from 750 participants, with 30 observations per image pair.
  - **priming task [PRIM]** for the same 3,000 image pairs of the IMG and DIS datasets. In a priming task, two stimuli (here: images) are presented in quick succession, and participants have to perform a task on the second image

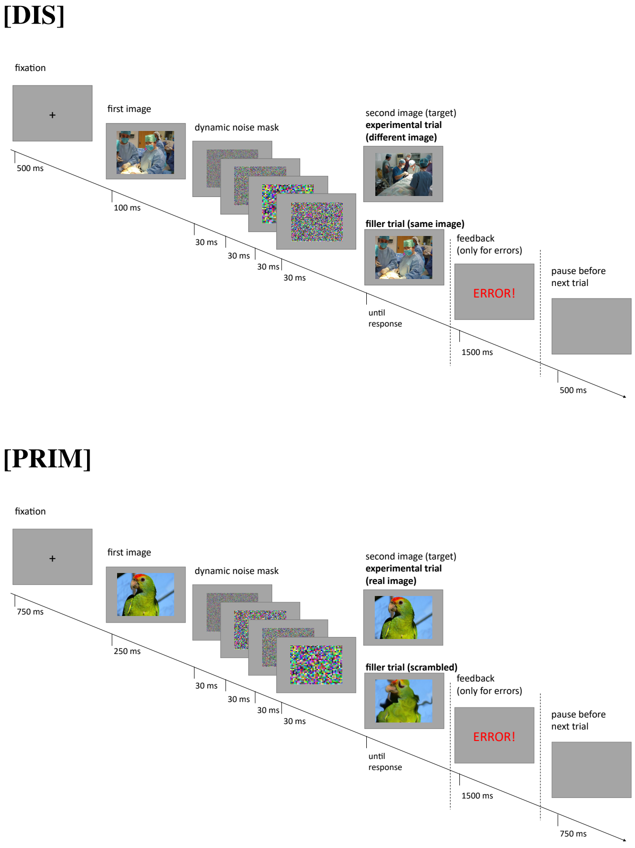

Figure 2: Schematic representations of experimental trials in the processing time paradigms. *Upper panel:* the discrimination task [DIS], in which participants have to decide whether the second image (the target) is identical to the first *lower panel:* priming task [PRIM], where participants have to decide whether the second image (the target) is a real image or a scrambled one. The behavioral variable of interest is the time until a response is made for the target.

only (here: judge whether a real or scrambled image has been presented by pressing one of two buttons); see Figure 2, lower panel for a schematic representation of an experimental trial. Responses are typically *faster* when the stimulus was preceded by a more visually similar stimulus, which primes (= facilitates processing of) the target. Data were collected from 750 participants, with 30 observations per image pair.

The target variable in these processing time studies is the mean response time for each image pair, after removing erroneous trials and outliers with far too slow or fast responses.

All datasets are publicly available in an OSF repository associated to the original study (Günther et al., 2023) at `https://doi.org/10.17605/OSF.IO/QVW9C`.

Vision Models

Models employed

For this study, we considered all pre-trained vision models available in the *MatConvNet* (Vedaldi & Lenc, 2015) and *Deep Learning Toolbox* (`https://github.com/matlab-deep-learning/MATLAB-Deep-Learning-Model-Hub`) packages for MATLAB. A full list of models is provided in Table 1.

General setup: Image and prototype representations

In line with previous studies (Battleday, Peterson, & Griffiths, 2020; Battleday et al., 2021; Günther et al., 2023; Petilli, Günther, Vergallito, Ciapparelli, & Marelli, 2021), we extracted the activation values in each convolutional and fully-connected layer of a model for a given input image (i.e., image embeddings) as representations for that image. In addition, we constructed prototype vectors for image labels as the centroid of 100–200 image embeddings of images tagged with that label (using the very same method presented in Günther et al., 2023; Petilli et al., 2021). For each image label, we obtain such a prototype representation for each layer of each considered model.

We used the cosine similarity metric to compute similarities between these image embeddings (at the same layer of the same model). In this manner, we can obtain a metric for the similarity between two individual images (for the IMG, DIS and PRIM datasets), the overall visual similarity between two categories denoted by their respective image labels (for the WORD dataset), and a typicality score as the similarity between an individual image embedding and the prototype vector for its category (for the TYP dataset). These metrics were computed for *each* layer of *each* model.

Results

Since relations between the model-derived similarities and the behavioral outcome variables are mostly non-linear (Günther et al., 2023), performance was assessed using Spearman rank correlations. All predictors (i.e., similarities based on each layer of each model) were ranked in terms of performance on each behavioral dataset, and these ranks were used to calculate three general-level evaluation metrics:

- The **rating performance** as the mean rank across the three rating datasets (IMG, WORD, and TYP)
- The **processing time performance** as the mean rank across the two processing-time datasets (DIS and PRIM)
- The **overall performance** as the mean rank across all behavioral datasets (compare Baroni et al., 2014; Gupta, Günther, Plag, Kallmeyer, & Conrad, 2021)

The results for the best-performing model layers for each evaluation metric are displayed in Table 2. We include the best-performing layer in the paper by Günther et al. (2023) (VGG-F, fully-connected layer 6) as a reference condition.

Table 1: Overview over the models investigated, including their number of layers, number of parameters, accuracy (measured as top-1 accuracy in the ImageNet classification task ILSVRC2012), and references to the papers introducing the models.

model	# layers	param. (mio.)	acc.	year	ref.
AlexNet	8	61.0	57.4	2012	Krizhevsky et al. (2012)
CaffeNet	8	61.0	57.4	2014	Jia et al. (2014)
DarkNet-19	19	20.8	74.0	2017	Redmon and Farhadi (2017)
DarkNet-53	53	41.6	76.5	2017	Redmon and Farhadi (2017)
DenseNet-201	201	20.0	75.9	2017	Huang, Liu, Van Der Maaten, and Weinberger (2017)
EfficientNet B0	82	5.3	74.7	2019	Tan and Le (2019)
GoogLeNet	22	7.0	66.3	2015	Szegedy et al. (2015)
Inception-ResNet-v2	164	55.9	79.6	2017	Szegedy et al. (2017)
Inception-v3	48	23.9	77.1	2016	Szegedy, Vanhoucke, Ioffe, Shlens, and Wojna (2016)
MobileNetV2	53	3.5	70.4	2018	Sandler, Howard, Zhu, Zhmoginov, and Chen (2018)
NASNet-Mobile	*	5.3	73.4	2018	Zoph, Vasudevan, Shlens, and Le (2018)
ResNet-18	18	11.7	69.5	2016	He, Zhang, Ren, and Sun (2016)
ResNet-50	50	25.6	74.5	2016	He et al. (2016)
ResNet-101	101	44.6	76.0	2016	He et al. (2016)
ResNet-152	152	60.3	77.0	2016	He et al. (2016)
ShuffleNet	50	1.4	63.7	2018	X. Zhang, Zhou, Lin, and Sun (2018)
SqueezeNet	18	1.2	55.2	2016	Iandola et al. (2016)
VGG-16	16	138.3	71.5	2014	Simonyan and Zisserman (2014)
VGG-19	19	143.7	71.3	2014	Simonyan and Zisserman (2014)
VGG-F	8	60.8	58.9	2014	Chatfield et al. (2014)
VGG-M	8	102.9	62.7	2014	Chatfield et al. (2014)
VGG-M-128	8	82.7	59.2	2014	Chatfield et al. (2014)
VGG-M-1024	8	87.2	62.2	2014	Chatfield et al. (2014)
VGG-M-2048	8	92.5	62.9	2014	Chatfield et al. (2014)
VGG-S	8	102.9	63.3	2014	Chatfield et al. (2014)
Xception	71	22.9	78.2	2017	Chollet (2017)

*NASNet-Mobile does not consist of a linear sequence of modules

Note that, for the PRIM dataset, participants tend to respond *faster* (that is, lower response times) if the two images are more similar; therefore, the target metric here is a *more negative* correlation.

As can be seen in Table 2, the overall best-performing representations (i.e., the model estimates most associated with behavioral variables) are provided by the GoogLeNet model, more specifically one of the representations in the 5th layer of the model (5a_3x3_reduce). These representations are also best-performing when it comes to predicting the arguably most fundamental types of behavioral data, similarity judgments within a given modality (i.e., between two different images [IMG] and between two different categories [WORD]).

Focusing only on the explicit rating data ([IMG], [WORD], and [TYP]), the best-performing representations are provided by the 7th layer (a fully-connected layer) of the VGG-M-1024 variant, closely followed by the same layer of the VGG-M-2048 variant. Although these perform slightly worse for the [IMG] dataset than the best-performing GoogLeNet layer (and very marginally worse for the [WORD] dataset), they make up for this with a near top-level performance in the [TYP] dataset (with the 50th convolutional layer of the DarkNet-53 model as the top-performer). However, these models fall behind a mean rank of 240 for the processing time data.

When focusing only on the processing time data ([DIS] and [PRIM]), the 6th layer (again, a fully-connected layer) of the VGG-M model (standard variant) performs best, with near top-level performance in both individual datasets (those top-performers being a layer of the EfficientNet B2 model and of the ResNet-50 model, respectively). However, conversely to the 7th layers in the VGG-M-1024 and VGG-M-2048 variants, these representations in turn fall behind for the rating data, with a mean rank of 157.

Comparison with model characteristics

In an additional step, we assessed the relation between the characteristics of a model (more specifically, their number of parameters and their top-1 classification accuracy (Russakovsky et al., 2015); see Table 1) and its performance on the behavioral datasets tested here. To this end, we equated the overall model performance with the performance of its best-performing layer, as measured by the mean rank.

We estimated two separate non-linear statistical models (GAMs; Wood, 2015; Fasiolo, Nedellec, Goude, & Wood, 2018), modelling mean rank (i.e., overall performance) as a function of model accuracy (deviance explained 37.2%, $r^2 = .295$, $p = .038$)[1] and number of parameters (deviance explained 11.6%, $r^2 = .051$, $p = .372$), the results of which are depicted in Figure 3. Note that a *lower* mean rank indicates better performance. As can be seen from these results and in these plots, medium levels of classification accuracy tend to be associated with better performance against behavioral data (with two local minima around 66% and around

[1] Model accuracy had no effect on rating performance ($p = .418$), or processing time performance ($p = .546$) individually.

Table 2: The best-performing layers across the different datasets, arranged by overall performance (the three top model layers in rows 1–3), rating performance (the two top model layers in rows 4–5), processing time performance (the top model layer in row 6), and performance in the individual datasets (the top model layers in row 1, as well as rows 7–10). The first number indicates the rank (ranging from a top value of 1 to a worst value of), the second number in brackets the Spearman correlation. The best-performing model layer in Günther et al. (2023) – VGG-F, layer fc6 – is listed as a baseline (row 10).
Note that *all* individual layers were considered for this analysis, not just the best-performing layer for each model.

row	model	layer	IMG	WORD	TYP	DIS	PRIM	rating	processing	overall
1	GoogLeNet	5a_3x3_reduce	**1 (0.774)**	**1 (0.666)**	35 (0.361)	65 (0.207)	88 (-0.088)	12.3	76.5	**38.0**
2	DarkNet-19	conv14	9 (0.740)	81 (0.646)	39 (0.359)	43 (0.211)	43 (-0.095)	43.0	43.0	43.0
3	GoogLeNet	5a_3x3	41 (0.707)	53 (0.649)	95 (0.339)	42 (0.211)	11 (-0.105)	63.0	26.5	48.4
4	VGG-M-1024	fc7	8 (0.741)	11 (0.660)	4 (0.389)	137 (0.199)	383 (-0.066)	**7.7**	260.0	108.6
5	VGG-M-2048	fc7	15 (0.734)	21 (0.657)	3 (0.392)	185 (0.192)	301 (-0.071)	13.0	243.0	105.0
6	VGG-M	fc6	36 (0.714)	288 (0.626)	147 (0.324)	12 (0.222)	8 (-0.107)	157.0	**10.0**	98.2
7	DarkNet-53	conv50	114 (0.648)	17 (0.658)	**1 (0.400)**	328 (0.176)	56 (-0.092)	44.0	192.0	103.2
8	EfficientNet B2	B12-D-conv2d-D*	47 (0.702)	42 (0.651)	160 (0.322)	**1 (0.231)**	129 (-0.083)	83	65	75.8
9	ResNet-50	Res5a-Branch2b	76 (0.679)	113 (0.643)	134 (0.327)	80 (0.205)	**1 (-0.128)**	107.7	40.5	80.8
10	VGG-F	fc6	24 (0.721)	244 (0.63)	176 (0.316)	7 (0.224)	18 (-0.102)	148	12.5	93.8

*layer *blocks-12-depthwise-conv2d-depthwise*

74%). We find no significant relation between the number of parameters and model performance (in terms of mean rank).

Discussion

Implications of the results

In the present study, we investigated which representations obtained from different computer-vision models best predict a battery of five large-scale behavioral datasets, including both rating data and processing time data. We find that the overall best-performing models are in fact quite "old" models, given the pace of the research cycle within the field: A layer of the GoogLeNet model (Szegedy et al., 2015) displays the overall highest performance across all five datasets, and different layers (of different variants) of the VGG-M model (Chatfield et al., 2014) display the best overall performance for the rating data and processing time data, respectively. Note that the differences in performance between the individual representations are meaningful and not trivial: For example, the difference in performance for the [IMG] dataset between the overall best GoogLeNet layer (0.774) and the overall second-best DarkNet-19 layer (0.740) is already more than three percentage points.

Over the last years, a lot of effort has gone into developing systems with ever better performance than these "older" models. With respect to the task these models are designed for – most prominently, image classification – this effort has reached impressive successes: As can be seen in Table 1, the top-1 accuracy for the ILSVRC 2012 validation data (Russakovsky et al., 2015) has increased dramatically, from around 60% in 2014/2015 to around 80%. In comparison, GoogLeNet (66.3%) and especially VGG-M (around 60% for all variants) definitely fall on the lower end of this scale. This however reveals an interesting rift opening with respect to model performance: Even though more recent models get better and better on their target tasks, this improvement in classification accuracy does not go along with improvements in predicting other types of data (in fact the contrary, compare Figure 3, upper panel). This is not to say that more recent models show low performance on this type of data: Representations from recent models and highly accurate models like DarkNet-19 (Redmon & Farhadi, 2017) are among the best-performing representations available. The critical point however still remains that the strong improvement in classification accuracy has not been accompanied by an *improvement* in predicting other types of data.

On the other hand, we find no clear connection between model complexity and top performance in the behavioral dataset: The GoogLeNet model is relatively small in terms of parameters (7 mio.) and comes with an intermediate number of layers (22), while the VGG-M models are quite large (around 90 to 100 mio. parameters) but have only a few layers (8). Therefore, one can neither conclude that a model needs to be very large and complex for top-level performance on behavioral data (consider especially the better performance of the VGG-M model vis-à-vis the conceptually and architecturally similar VGG-16 and VGG-19 models), nor that it needs to be particularly small and efficient (compare also Figure 3, lower panel).

At this point, we can only speculate *why* more recent and more classification-accurate models don't perform better for behavioral data. One explanation may be that the models are optimized to predict a very *specific* image class, and only ex-

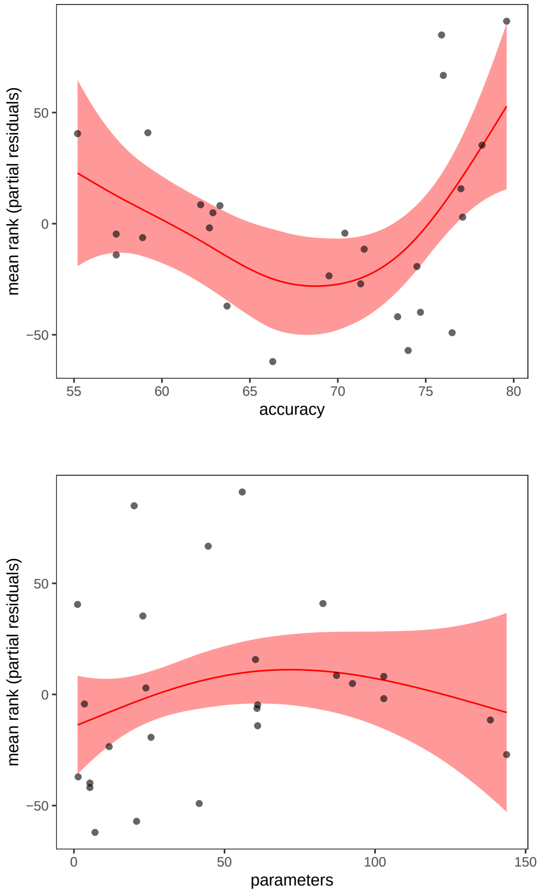

Figure 3: The relation between a model's accuracy (upper panel) and number of parameters (lower panel) on the performance across all beahvioral datasets (measured as mean rank; the graphs show the partial residuals of a GAM analysis of this outcome variable). Each individual data point represents the best-performing layer of one of the models tested here.

act matches as a hit when calculating accuracy – with the misclassification of a *spotted salamander* as a *European fire salamander* treated as a miss in the same way a mis-classification as a *toaster* is. This may lead the models to weight relatively specific details to a similar or maybe even larger degree than the overall structure/"gestalt" of the depicted object. Human judgments and responses, on the other hand, are more driven by these general-level similarities (e.g. Hebart et al., 2020) rather than details (even if those are very informative for classification); this might lead to the observed discrepancy between classification accuracy and performance on behavioral datasets. However, we want to stress again that this is speculation on an open question, and more research is necessary to properly investigate and explain this discrepancy.

Limitations and future directions

At this point, we need to emphasize that all the issues discussed so far are based on the results of our evaluation, and therefore necessarily restricted to the models analysed here. However, there may well be models we did not consider here which contradict our findings (i.e., a high-accuracy model that simultaneously has a higher performance on behavioral data than the best-performing models identified here). In fact, in the context of successful transfer learning, we would consider this highly desirable, and hope that our study can give an impetus to systematically include also behavioral data in the search for an overall well-performing model. While one may dismiss the behavioral data analysed here as not relevant for evaluating the performance of computer vision models, we argue that at the very least recognizing which images are more or less similar to one another should be considered one of the core prerequisites for a general-level vision model, analogous to semantic models predicting semantic similarity and relatedness data in the NLP domain (Baroni et al., 2014; Bruni et al., 2014; Finkelstein et al., 2001; Hill et al., 2015).

In general, a desirable direction for future work in the field would be to develop general-level models that do not only excel in one particular task, but perform well across a range of different tasks (including but not limited to behavioral data). Ideally, in the spirit of successful transfer learning, this would not simply mean *optimizing* a single model for a range of different tasks, but instead *testing* such a model on a battery of tasks it was not optimized for (Srivastava et al., 2022). Following up on our suspicion that the lack of improvement in representation learning could be the result of hyper-engineering to distinguish very specific (and somewhat arbitrary) categories, we speculate that possible routes of advancement to achieve models representations that better capture a general similarity structure could be as follows: On the one hand, the training objective of the models could be altered to not only consider *exact* hits among a set of candidate categories, but to also partially reward *close* hits, for example based on their word embedding similarity or their WordNet distance to the correct target (thus rewarding the classification of a *poodle* as a *dalmatian* or as a *dog* more than as a *Persian cat*, and that more than as a *pillow*; see also De Deyne, Navarro, Collell, & Perfors, 2021). On the other hand, the training sets of the models could be altered to more closely approximate human visual experience rather than over-representing certain categories (Elgendy, 2020), or to include more than one correct label per image (Silberer, Zarieß, & Boleda, 2020).

We argue that such developments would be interesting from an engineering/transfer learning viewpoint (since a successful general-level model could be applied to new tasks that it was not originally optimized for), but also for the application of such systems as models of human visual representations in cognitive (neuro)science.

Data availability

Data and the analysis script for this study are available at `https://osf.io/sx5u3`.

Acknowledgments

This work was supported by a Research Grant (Project Number 452322374) from the German Research Foundation (DFG), awarded to Fritz Günther.

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL 2014* (pp. 238–247). East Stroudsburg, PA: ACL.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature Communications*, *11*, 5418.

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2021). From convolutional neural networks to models of higher-level cognition (and back again). *Annals of the New York Academy of Sciences*, *1505*, 55-78.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1-47.

Byerly, A., Kalganova, T., & Ott, R. (2022). The current state of the art in deep learning for image classification: A review. In K. Arai (Ed.), *Intelligent computing* (pp. 88–105). Cham: Springer International Publishing.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint; arXiv:1405.3531*.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1251–1258).

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*, 305-317.

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*, e12922.

Elgendy, M. (2020). *Deep learning for vision systems*. Shelter Island, NY: Manning.

Fasiolo, M., Nedellec, R., Goude, Y., & Wood, S. N. (2018). Scalable visualisation methods for modern Generalized Additive Models. *Arxiv preprint*. Retrieved from `https://arxiv.org/abs/1707.03307`

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*, 116–131.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Representation learning. In *Deep learning* (pp. 524–554). MIT press.

Günther, F., Marelli, M., Tureski, S., & Petilli, M. A. (2023). Vispa (vision spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. *Psychological Review*, *130*, 896–934.

Gupta, A., Günther, F., Plag, I., Kallmeyer, L., & Conrad, S. (2021). Combining text and vision in compound semantics: Towards a cognitively plausible multimodal model. In *Proceedings of the 17th conference on natural language processing (konvens 2021)* (pp. 218–222).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr 2016)* (pp. 770–778).

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*, 1173–1185.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*, 665–695.

Hollis, G. (2018). Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior Research Methods*, *50*, 711–729.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr2017)* (pp. 4700–4708).

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, *8*, 1726.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105).

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*, 2017–2031.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*, 2648–2669.

Petilli, M. A., Günther, F., Vergallito, A., Ciapparelli, M., & Marelli, M. (2021). Data-driven computational models reveal perceptual simulation in word comprehension. *Journal of Memory and Language*, *117*, 104194.

Redmon, J., & Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr 2019)* (pp. 7263–7271).

Roads, B. D., & Love, B. C. (2021). Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3547–3557).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … others (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr 2018* (pp. 4510–4520).

Silberer, C., Zarieß, S., & Boleda, G. (2020). Object Naming in Language and Vision: A Survey and a New Dataset. In *Proceedings of the 12th international conference on language resources and evaluation (lrec 2020).* Retrieved from `https://aclanthology.org/2020.lrec-1.710/`

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint, arXiv:1409.1556*.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., … others (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 31).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition.*

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2818–2826).

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).

Vedaldi, A., & Lenc, K. (2015). Matconvnet: Convolutional neural networks for Matlab. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 689–692).

Wood, S. N. (2015). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=mgcv` (R package version 1.8-7)

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 586–595).

Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr 2018)* (pp. 6848–6856).

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8697–8710).