CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 136 **July 2, 2022**

System identification of neural systems: If we got it right, would we know?

Yena Han, Tomaso Poggio, Brian Cheung

Abstract

Various artificial neural networks developed by engineers have been evaluated as models of the brain, such as the ventral stream in the primate visual cortex. After being trained on large datasets, the network outputs are compared to recordings of biological neurons. Good performance in reproducing neural responses is taken as validation for the model. This system identification approach is different from the traditional ways to test theories and associated models in the natural sciences. Furthermore, it lacks a clear foundation in terms of theory and empirical validation. Here we begin characterizing some of these emerging approaches: what do they tell us? To address this question, we benchmark their ability to correctly identify a model by replacing the brain recordings with recordings from a known ground truth model. We evaluate commonly used identification techniques such as neural regression (linear regression on a population of model units) and centered kernel alignment (CKA). Even in the setting where the correct model is among the candidates, we find that the performance of these approaches at system identification is quite variable; it also depends significantly on factors independent of the ground truth architecture, such as scoring function and dataset.

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

System identification of neural systems: If we got it right, would we know?

Yena Han
Center for Brains, Minds and Machines
Massachusetts Institute of Technology

Tomaso Poggio
Center for Brains, Minds and Machines
Massachusetts Institute of Technology

Brian Cheung
Dept. of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Various artificial neural networks developed by engineers have been evaluated as models of the brain, such as the ventral stream in the primate visual cortex. After being trained on large datasets, the network outputs are compared to recordings of biological neurons. Good performance in reproducing neural responses is taken as validation for the model. This system identification approach is different from the traditional ways to test theories and associated models in the natural sciences. Furthermore, it lacks a clear foundation in terms of theory and empirical validation. Here we begin characterizing some of these emerging approaches: what do they tell us? To address this question, we benchmark their ability to correctly identify a model by replacing the brain recordings with recordings from a known ground truth model. We evaluate commonly used identification techniques such as neural regression (linear regression on a population of model units) and centered kernel alignment (CKA). Even in the setting where the correct model is among the candidates, we find that the performance of these approaches at system identification is quite variable; it also depends significantly on factors independent of the ground truth architecture, such as scoring function and dataset.

1 Introduction

Over the last two decades, the dominant approach for machine learning engineers in search of better performance has been to use standard benchmarks to rank networks from most relevant to least relevant. This practice has driven much of the progress in the machine learning community. A standard comparison benchmark enables the broad validation of successful ideas. Recently such benchmarks have found their way into neuroscience with the advent of experimental frameworks like Brain-Score [23], and Algonauts [3], where artificial models compete to predict recordings from real neurons in animal brains. Can engineering approaches like this be helpful in the natural sciences?

Understanding natural intelligence at the level of the underlying neural circuits requires developing model systems that reproduce the abilities of their biological analogs while respecting all the constraints provided by biology, including anatomy and biophysics. However, the "engineering approach" described above ranks models that predict neural responses better to be better models of animal brains. While such absolute rankings may be a good measure of absolute performance in approximating the neural responses, it is an open question whether they lead to better models of the underlying brain. It is difficult to account for how well different models respect anatomical and physiological data in the ranking. In this paper, we argue that even in the most favorable settings,

Preprint. Under review.

ranking only based on correlation with neural data does not guarantee identifying the correct model architecture amongst incorrect ones.

We find several culprits for this lack of identification. Two examples are *dataset* and *scoring function*. We evaluate the factors contributing to similarity measures of candidate neural network models to biological systems through a series of simulated experiments with known ground truth. This is particularly important when identifying drastically different architectural motifs of the target system, such as convolution vs. attention.

1.1 System identification from leaderboards

The approach described by [30] involves first taking a candidate model trained to perform a biologically relevant task and then measuring the regression performance of this model to a biological system. Multiple candidates are compared, and models with the leading regression score are deemed closest to biology. As improvements to regression scores are made over time, ideally, more biologically relevant candidates emerge. Nevertheless, consider the following thought experiment:

Two artificial models with a distinctly different architectures are trained on the same data and happen to be similar in reproducing neural activities (target model). At initialization, the two models had differing similarities to the biological system.

Based on this result, it is difficult to conclude whether the differences between these two models are driven by biologically relevant motifs from the architecture, biological relevance of the training process, or a combination of the two. Such ambiguity is due to the many-to-one mapping of a model onto a leaderboard score. Multiple factors can lead to a similar score, and our work shows that these factors play a role in many standard artificial models compared to biological systems.

1.2 Data impacts identifiability

Candidate models often used for comparison are first trained to converge on a machine learning dataset. The learned representations of these models are then used to predict neural/cognitive recordings of a biological system on a stimulus dataset. The machine learning dataset significantly influences the representations generated by these highly parameterized models [34]. In turn, this can have a significant influence on downstream identifiability. We investigate how the stimulus datasets can impact identifiability when drawn from different sources closer to or farther away from the original machine learning task of the candidate model.

2 Related Work

As modern neural network models have grown larger in unison with the corresponding resources to train these models, pre-trained reference models have become more widely available in research [29]. Consequently, the need to compare these references along different metrics has followed suit. [15] explored using different similarity measures between the layers of artificial neural network models. [15] propose various properties a similarity measure should be invariant such as orthogonal transformations and isotropic scaling while not invariant to invertible linear transformations. [15] found Centered Kernel Alignment (CKA), a method very similar to Representation Similarity Matrices [16], to best satisfy these requirements. [6] explored the sensitivity of methods like canonical correlation analysis, centered kernel alignment, and orthogonal procrustes distance to changes in factors that do not impact the functional behavior of neural network models.

3 Background and Methods

The two predominant approaches to evaluating computational models of the brain are using metrics based on single-unit response predictivity and population-level representational similarity. The first measures how well a model can directly predict the activations of individual units, whereas the second metric measures how correlated the variance of internal representations are. Consistent with the typical approaches, we study the following neural predictivity scores: Neural Regression and Centered Kernel Alignment (CKA).

In computational neuroscience, we usually have a neural system (brain) that we are interested in modeling. We call this network a *target* and the proposed candidate model a *source*. Formally, for a layer with p_1 units in a source model, let $X \in \mathbb{R}^{n \times p_1}$ be the matrix of representations with p_1 features over n stimulus images. Similarly, let $Y \in \mathbb{R}^{n \times p_2}$ be a matrix of representations with p_2 features of the target model (or layer) on the same n stimulus images.

3.1 Neural Regression

Closely following the procedure developed by previous works [23, 31, 4], we linearly project the feature space of a single layer in a source model to map onto a single unit in a target model (a column of Y). The neural regression score is the Pearson's correlation $r(\cdot, \cdot)$ coefficient between the predicted responses of a source model and the ground-truth target responses to a set of probe images from a stimulus dataset.

$$\hat{\beta} = \text{argmin}_\beta ||Y - XS\beta||_F^2 + \lambda ||\beta||_F^2 \tag{1}$$

$$NR(X, Y) = r(XS\hat{\beta}, Y) \tag{2}$$

We first extract activations on the same set of probe images for source and target models. To reduce computational costs without sacrificing predictivity, we apply sparse random projection, $S \in \mathbb{R}^{p_1 \times q_1}$ for $q_1 << p_1$, on the activations of the source model [4]. This projection reduces the dimensionality of the features to q_1 while still preserving relative distances between points [19]. We apply ridge regression on every layer of a source model to predict a target unit using these features. We use 90% of the probe images for linear fitting and test on 10%, cross-validated 10 times. As there are multiple target units, the median of Pearson's correlation coefficients between predicted and true responses is the aggregate score for layer-wise comparison between source and target models. Note that a layer of a target model is usually assumed to correspond to a visual area, e.g. V1 or IT, in the visual cortex. For a layer-mapped model, we report maximum neural regression scores across source layers for target layers.

3.2 Centered Kernel Alignment

Another widely used type of metric builds upon the idea of measuring the representational similarity between the activations of two neural networks for each pair of images. While variants of this metric abound, including RSA or re-weighted RSA [16, 14], we use CKA as [15] showed strong correspondence between layers of models trained with different initializations, which we will further discuss as a validity test we perform. We consider linear CKA in this work:

$$\text{CKA}(X, Y) = \frac{||Y^T X||_F^2}{||X^T X||_F ||Y^T Y||_F} \tag{3}$$

We also call CKA a neural predictivity score because a target network is observable, whereas a source network gives predicted responses.

3.3 Identifiability Index

To quantify how selective neural predictivity scores are when a source matches the target architecture compared to when the architecture differs between source and target networks, we define an identifiability index as:

$$\text{Identifiability Index} = \frac{\text{Score}(\text{source} = \text{target}) - \text{Mean Score}(\text{source} \neq \text{target})}{\text{Score}(\text{source} = \text{target}) + \text{Mean Score}(\text{source} \neq \text{target})} \tag{4}$$

Here, the index is selectivity to two identical architectures. Previous works [7, 10] defined selectivity indices in the same way in similar contexts, such as the selectivity of a neuron to specific tasks.

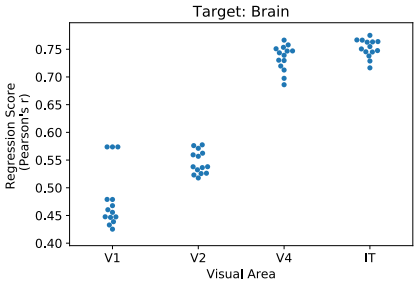


Figure 1: Neural regression scores of deep neural networks for brain activations in the macaque visual cortex. For V1, the top performing three models are in VOneNet family [5], which are explicitly designed to mimic the known properties of V1.

3.4 Simulated Environment

If a target network is a brain, it is essentially a black box, making it challenging to understand the properties or limitations of the comparison metrics. Therefore, we instead use artificial neural networks of our choice as targets for our experiments.

We investigate the reliability of a metric to compare models, mainly to discriminate the underlying computations specified by the model's architecture. We intentionally create favorable conditions for identifiability in a simulated environment where the ground truth model is a candidate among the source models. Taking these ideal conditions further, our target and source models are deterministic and do not include adverse conditions typically encountered in biological recordings, such as noise and temporal processing. We consider the following architectures:

Convolutional Networks: AlexNet [17], VGG11 [26], ResNet18 [12]

Transformer Networks: ViT-B/32 [8]

Mixer Networks: MLP-Mixer-B/16 [27]

These architectures are emblematic of the vision-based models used today. Each architecture also has a distinct motif, making it unique from other models. For example, transformer networks use the soft-attention operation as a core motif, whereas convolutional networks use convolution. Moreover, mixer networks [27, 28, 21] uniquely perform fully-connected operations over image patches, alternating between the feature and patch dimensions.

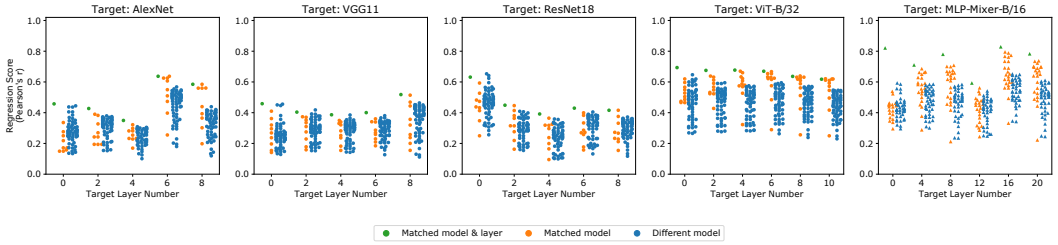
4 Results

4.1 Different models trained on a large-scale dataset reach equivalent neural predictivity

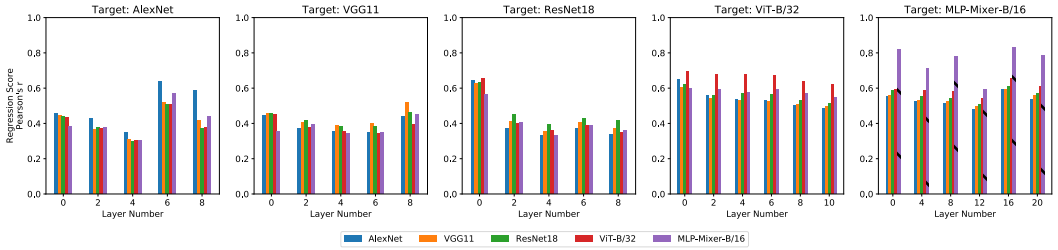
We compare various artificial neural networks with publicly shared neural recordings in primates [20, 9] via Brain-Score framework [23]. Our experiments show that the differences between markedly different neural network architectures are minimal after training (Figure 1), consistent with the previous work [23, 18, 4]. Previous works focused on the relative ranking of models and investigated which model yields the highest score. However, if we take a closer look at the result, the performance difference is minimal, with the range of scores having a standard deviation < 0.03 (for V2=0.021, V4=0.023, IT=0.016) except for V1. For V1, VOneNets [5], which explicitly build in properties observed from experimental works in neuroscience, significantly outperform other models. Notably, the models we consider have quite different architectures based on different combinations of various components, such as convolutional layers, attention layers, and skip connections. This suggests that architectures with different computational operations reach almost equivalent performance after training on the same large-scale dataset, i.e., ImageNet.

4.2 Identification of architectures in an ideal setting

One potential interpretation of the result would be that different neural network architectures are indeed equally good (or bad) models of the visual cortex. An alternative explanation would be



(a) Neural regression scores for every pair of layers



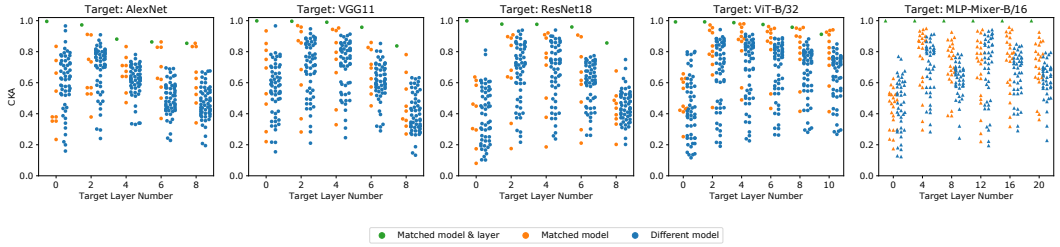
(b) Neural regression scores for top-performing layer in each source network

Figure 2: Neural regression scores for artificial neural networks. (a): Each data point indicates the neural regression score of a layer in a source network in predicting activations for a layer in a target network. The case of an ideal correct mapping is when both architecture type and layer match between source and target networks, shown in green. (b): Aggregate top scores, which are the top predictivity scores for each source model against a target layer, are shown. The ranking of source models is typically decided based on these aggregate scores [23]. For both (a) and (b), we use different initialization seeds for source networks of the same architecture type as the target, except for MLP-Mixer-B/16, for which we test identical weights, the most ideal setting. We show the results for MLP-Mixer-B/16 in triangular markers and bar plots with a pattern to indicate the difference from other targets.

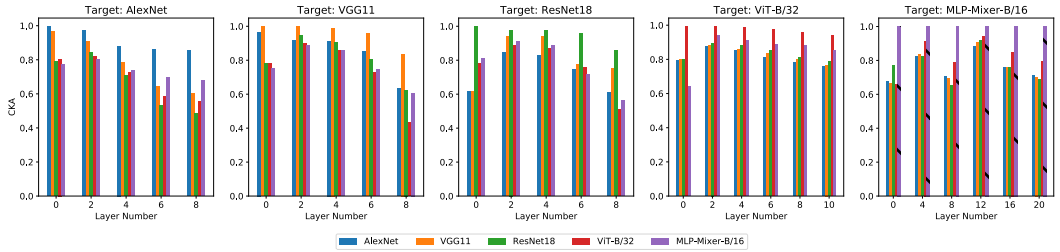
that the method we use to compare models with the brain has limitations in identifying the precise computational operation. To test the hypothesis, we turn our attention to the case where underlying target neural networks are known instead of being a black box as with biological brains. Specifically, we replace the target neural recordings with activations of an artificial neural network. By examining whether the candidate source model with the highest predictivity is identical to the target model, we can evaluate to what extent we can identify architectures with the current approach.

We first compare various source models (AlexNet, VGG11, ResNet18, ViT-B/32, MLP-Mixer-B/16) with a target network, the same architecture as one of the source models and is trained on the same dataset but initialized with a different seed. We test images of synthetic objects studied in [20] to be consistent with the evaluation pipeline of Brain-Score. The ground-truth source model will yield a higher score than other models if the model comparison pipeline is reliable. For most target layers, except for those in VGG11, source layers with the highest score are the matching layers in the same network (Figure 2). However, strikingly, for early and intermediate layers of target VGG11, the best-matched layers belong to a source model that is not VGG11. The correct source model outperforms other models only for three layers (two of which are the last two layers when close to the final classification task). The first layer of ResNet18 is also predicted best by ViT-B/32. In other words, given the activations of VGG11, for instance, and based on neural regression scores, we would make an incorrect prediction that the system's underlying architecture is closest to a ResNet18.

In addition, because of our ideal setting, where an identical network is one of the source models, we expect to see a significant difference between matching and non-matching models. However, for some target layers in AlexNet and ResNet18, although the layer with the highest score may be the matching layer in the same architecture, neural regression scores for other source models do not show a significant decrease in predictivity. Taken together, this result suggests that the identification of the underlying architectures of unknown neural systems is far from perfect.



(a) CKA scores for every pair of layers



(b) CKA scores for top-performing layer in each source network

Figure 3: CKA for different source and target networks. The experimental setup is identical to Figure 2 besides using CKA instead of neural regression. As in Figure 2, when we test MLP-Mixer-B/16 as a target and the source network type matches the target, weights are identical; thus CKA is trivially 1.

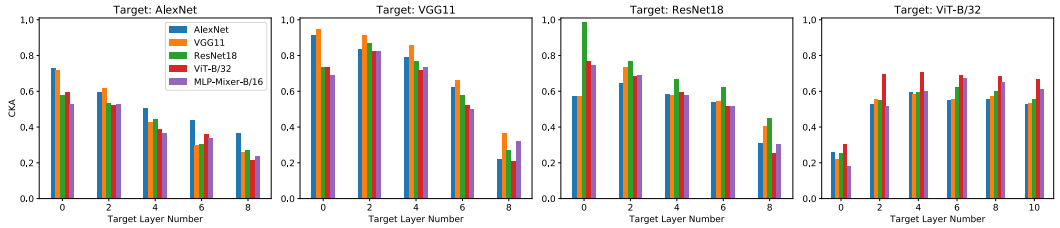


Figure 4: CKA scores when only a subset (1%) of units in a target model are available to be recorded. The constraint is tested to examine whether CKA is reliable in a setting closer to a biological experiment.

4.3 Using CKA to identify architectures

Next, we study if CKA can be an alternative measure to neural regression for system identification. Again, we compare different source models to a target model, also an artificial neural network. For all layers of the target models we tested, the ground-truth source model achieved the highest score with a significant margin (Figure 3).

When applying CKA to compare representations, one assumption is that all units in the models are available to measure. This availability may be a strong assumption, as neurons in the order of hundreds are usually recorded in animal brains in neuroscience. We vary the ratio of target units included to examine how robust CKA is when only a randomly sampled subset of target neurons are available. When 1% of target units are available for every layer of a target network, we show that the correct source model can still be identified for most layers, but start to observe some layers that are either incorrect or have similar scores across models (Figure 4). Overall, the degree of identifiability decreases, and we expect settings that are more consistent with biology will have even more constraints and noise.

4.4 Effects of the distribution of probe images on identifiability

A potentially significant variable that is overlooked in comparing computational models of the brain is the type of probe images. What types of probe images are suited for evaluating competing models?

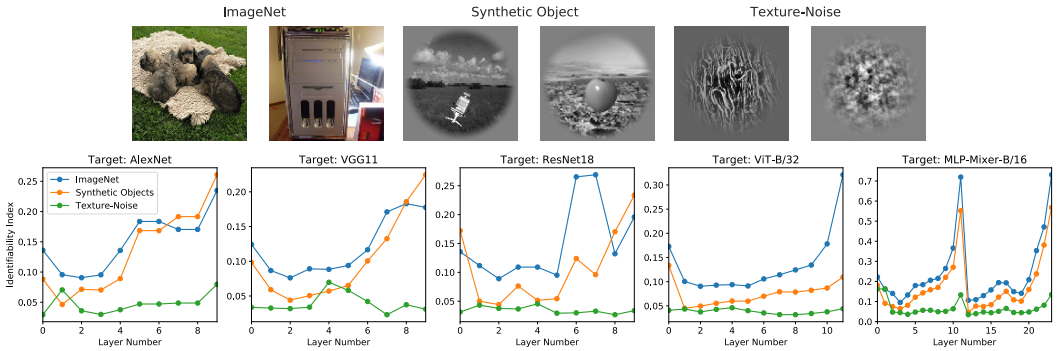


Figure 5: **Top** Sample images of each probe image type. **Bottom** Architectural identifiability index using CKA for different types of probe images and target networks.

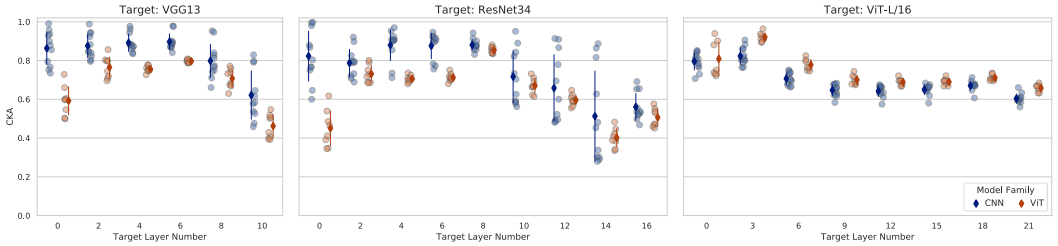


Figure 6: CNNs and ViTs of different architectural variants are compared with two CNNs and a ViT target networks. Each datapoint is the maximum CKA of an architecture for corresponding target layers. Markers with darker shades indicate mean CKA of the corresponding model class, and error bars are standard deviation.

In Brain-Score, probe images for comparing models of the high-level visual areas, V4 and IT, are images of synthetic objects [20]. In contrast, those for the lower visual areas, V1 and V2, are images of texture and noise [9]. To examine the effect of using different probe images, we test images of synthetic objects, texture and noise, and ImageNet, which are more natural images than the first two datasets.

In Figure 5, we analyze Identifiability Index for different probe images. More natural probe images (i.e., synthetic objects and ImageNet) show higher identifiability than texture and noise images for all target models. Notably, even for early layers in target models, which would correspond to V1 and V2 in the visual cortex, texture and noise images fail to give higher identifiability. Also, between images of synthetic objects and ImageNet, ImageNet shows higher identifiability. Overall, our results suggest that natural probe images with more variability help identify the architecture.

It is important to note that the images of texture and noise we use in the experiment indeed help characterize certain aspects of V1 and V2 in the previous work [9]. More specifically, the original work investigated the functional role of V2 in comparison with V1 by showing that naturalistic structure modulates V2. Although the image set plays as an effective variable in a carefully designed experiment for a more targeted hypothesis, it does not translate as a sufficient test set for any hypothesis, such as evaluating different neural networks.

4.5 Challenges of identifying key architectural motifs with CKA

Interesting hypotheses for a more biologically plausible design principle of brain-like models often involve key high-level architectural motifs. For instance, potential questions are whether recurrent connections are crucial in visual processing or, with the recent success of transformer models in deep learning, whether the brain similarly implements computations like attention layers in transformers. The details beyond the key motif, such as the number of layers or exact type of activation functions, may vary and be underdetermined within the scope of such research questions. Likewise, it is unlikely that candidate models proposed by scientists align with the brain at every level, from low-level

specifics to high-level computation. Therefore, an ideal methodology for comparing models should help separate the key properties of interest while being invariant to other confounds.

Considering it is a timely question, with the increased interests in transformers as models of the brain in different domains [24, 1], we focus on the problem of identifying convolution vs. attention. We test 12 Convolutional Networks and 8 Vision Transformers of different architectures, and to maximize identifiability, we use ImageNet probe images. Note that an identical architecture with the target network is not included as a source network. Overall, Figure 6 shows that mean CKA is higher when there are correspondences between the target and source model classes than when there are not. However, one layer in VGG13 (layer 8), 6 layers in ResNet34 (layer 2, 8-16), and one layer in ViT-L/16 (layer 0) do not show a statistically significant difference between the two model classes based on Welch's t-test with $p < 0.01$ used as a threshold.

Furthermore, we expect the inter-class variance for ViTs is under-estimated, given that tested ViTs are limited to two previous works [8, 33]. The significant variance among source models suggests that model class identification can be incorrect depending on the precise variation we choose, especially if we rely on a limited set of models. A quick but essential remedy for this issue is to include wide-ranging variants of a model class rather than to test a single model before concluding high-level key computations.

5 Conclusion

Under idealized settings, we tested the identifiability of various artificial neural networks with differing architectures. Our results indicate that identifiability among these models is highly variable and dependent on the properties of the target architecture and the stimulus data used to probe the candidate models. When the stimulus dataset is closer to the training domain of the target model, identifiability improves.

While system identification depends on many factors which can lead to variable scores, both neural regression and CKA give reasonable identification capability under unrealistically ideal conditions. However, we find metrics like CKA have slightly better reliability than neural regression.

6 Future Work and Broader Impacts

Neural networks have been proven [31] to be able to reproduce neural data. This is good but not too surprising because linear regression on random features can yield universal approximation [22, 32]. From the point of view of a neuroscientist, the problem is that several rather different architectures seem to be able to account rather well for the neural data, with differences among them that may be too small or too inconsistent to be of sufficiently clear significance.

At least since the Neocognitron [11] and HMAX [25], networks that can readily learn from data have led to models which can better explain neuroscience data. However, the ability to learn from data can potentially become a confound, obscuring well-understood underlying factors that lead to accurate identification of biological systems. It is worthwhile to think about the history of models in neuroscience to appreciate some of the potential risks of these new approaches. Until now most of the biological models were well motivated by various aspects of the underlying neuroscience, usually spanning several different experimental methods such as anatomy, physiology and biophysics. Often, however, models that originate in the engineering community are not further constrained by neuroscience and are then tested mostly on their ability to reproduce neural activities. An interesting example is offered by [2], which compares many different models with respect to their ability of reproducing neural responses in IT to face images and conclude that the *2D morphable model* is best. Such a model, originally proposed for faces by [13], happens to be the least biologically plausible among all the models considered, requiring operations such as correspondence and vectorization that do not have an obvious biological implementation in terms of neurons and synapses. Ranking models this way is arguably only a first step in identifying the underlying neural architecture. As the overlap between the neuroscience and machine learning communities continues to grow, next steps should investigate the underlying causes these rankings.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

[1] William Berrios and Arturo Deza. "Joint rotational invariance and adversarial training of a dual-stream Transformer yields state of the art Brain-Score for Area V4". In: *arXiv preprint arXiv:2203.06649* (2022).

[2] Le Chang et al. "Explaining face representation in the primate brain using different computational models". In: *Current Biology* 31.13 (2021), 2785–2795.e4. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2021.04.014. URL: https://www.sciencedirect.com/science/article/pii/S0960982221005273.

[3] Radoslaw Martin Cichy et al. "The Algonauts Project 2021 Challenge: How the Human Brain Makes Sense of a World in Motion". In: *arXiv preprint arXiv:2104.13714* (2021).

[4] Colin Conwell et al. "Neural Regression, Representational Similarity, Model Zoology & Neural Taskonomy at Scale in Rodent Visual Cortex". In: *Advances in Neural Information Processing Systems* 34 (2021).

[5] Joel Dapello et al. "Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 13073–13087.

[6] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. "Grounding Representation Similarity with Statistical Testing". In: *arXiv preprint arXiv:2108.01661* (2021).

[7] Katharina Dobs et al. "Brain-like functional specialization emerges spontaneously in deep neural networks". In: *Science advances* 8.11 (2022), eabl8913.

[8] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[9] Jeremy Freeman et al. "A functional and perceptual signature of the second visual area in primates". In: *Nature neuroscience* 16.7 (2013), pp. 974–981.

[10] Winrich A Freiwald and Doris Y Tsao. "Functional compartmentalization and viewpoint generalization within the macaque face-processing system". In: *Science* 330.6005 (2010), pp. 845–851.

[11] K. Fukushima. "Neocognitron: A self-organizing neural network for a mechanism of pattern recognition unaffected by shift in position". In: *Biological Cybernetics* 36.4 (1980), pp. 193–202.

[12] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[13] Michael J. Jones and Tomaso A. Poggio. "Multidimensional morphable models". In: *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)* (1998), pp. 683–688.

[14] Seyed-Mahdi Khaligh-Razavi et al. "Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models". In: *Journal of Mathematical Psychology* 76 (2017), pp. 184–197.

[15] Simon Kornblith et al. "Similarity of neural network representations revisited". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3519–3529.

[16] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. "Representational similarity analysis-connecting the branches of systems neuroscience". In: *Frontiers in systems neuroscience* 2 (2008), p. 4.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[18] Jonas Kubilius et al. "Brain-like object recognition with high-performing shallow recurrent ANNs". In: *Advances in neural information processing systems* 32 (2019).

[19] Ping Li, Trevor J Hastie, and Kenneth W Church. "Very sparse random projections". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, pp. 287–296.

[20] Najib J Majaj et al. "Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance". In: *Journal of Neuroscience* 35.39 (2015), pp. 13402–13418.

[21] Luke Melas-Kyriazi. "Do you even need attention? a stack of feed-forward layers does surprisingly well on imagenet". In: *arXiv preprint arXiv:2105.02723* (2021).

[22] Ali Rahimi and Benjamin Recht. "Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller et al. Vol. 21. Curran Associates, Inc., 2008. URL: https://proceedings.neurips.cc/paper/2008/file/0efe32849d230d7f53049ddc4a4b0c60-Paper.pdf.

[23] Martin Schrimpf et al. "Brain-score: Which artificial neural network for object recognition is most brain-like?" In: *BioRxiv* (2020), p. 407007.

[24] Martin Schrimpf et al. "The neural architecture of language: Integrative modeling converges on predictive processing". In: *Proceedings of the National Academy of Sciences* 118.45 (2021).

[25] Thomas Serre, Aude Oliva, and Tomaso Poggio. "A feedforward architecture accounts for rapid categorization". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.15 (2007), pp. 6424–6429. ISSN: 0027-8424. URL: http://cat.inist.fr/?aModele=afficheN%5C&cpsidt=18713198.

[26] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[27] Ilya O Tolstikhin et al. "Mlp-mixer: An all-mlp architecture for vision". In: *Advances in Neural Information Processing Systems* 34 (2021).

[28] Hugo Touvron et al. "Resmlp: Feedforward networks for image classification with data-efficient training". In: *arXiv preprint arXiv:2105.03404* (2021).

[29] Ross Wightman. *PyTorch Image Models*. https://github.com/rwightman/pytorch-image-models. 2019. DOI: 10.5281/zenodo.4414861.

[30] Daniel LK Yamins and James J DiCarlo. "Using goal-driven deep learning models to understand sensory cortex". In: *Nature neuroscience* 19.3 (2016), pp. 356–365.

[31] Daniel LK Yamins et al. "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the national academy of sciences* 111.23 (2014), pp. 8619–8624.

[32] Gilad Yehudai and Ohad Shamir. "On the Power and Limitations of Random Features for Understanding Neural Networks". In: *CoRR* abs/1904.00687 (2019). arXiv: 1904.00687. URL: http://arxiv.org/abs/1904.00687.

[33] Li Yuan et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 558–567.

[34] Chengxu Zhuang et al. "Unsupervised neural network models of the ventral visual stream". In: *Proceedings of the National Academy of Sciences* 118.3 (2021).

A Appendix

A.1 Model details for Section 4.1: Brain-Score

Below is the full list of models tested on the benchmarks of Brain-Score as reported in Section 4.1. In addition to testing vision models pre-trained on ImageNet available from PyTorch's torchvision model package version 0.12, we test VOneNets that are pre-trained on ImageNet and made publicly available by the authors [5]. VOneNets are also a family of CNNs.

Convolutional Networks: AlexNet, VGG11, VGG13, VGG19, ResNet18, ResNet34, ResNet50, ResNet101, VOneAlexNet, VOneResNet50, VOnetCORnet-S

Transformer Networks: ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32

A.2 Model details for Section 4.5: finding the key architectural motif

For each target network reported in Section 4.5, namely VGG13, ResNet34, and ViT-L/16, below is the full list of source models tested to compare two model classes, CNN and transformer. For Tokens-to-token ViTs (T2T) [33], we use models pre-trained on ImageNet and released by the authors. All other models are also pre-trained on ImageNet, available from PyTorch's torchvision model package version 0.12.

Convolutional Networks: AlexNet, VGG11, VGG13, VGG16, VGG13_bn, ResNet18, ResNet34, ResNet50, Wide-ResNet50_2, SqueezeNet1_0, Densenet121, MobileNet_v2

Transfomer Networks: ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/32, T2T-ViT_t-14, T2T-ViT_t-19, T2T-ViT-7, T2T-ViT-10